The Dissertation Committee for Eric Scott Hersh certifies that this is the approved version of the following dissertation:

THE LONG TAIL OF HYDROINFORMATICS: IMPLEMENTING BIOLOGICAL AND OCEANOGRAPHIC INFORMATION IN HYDROLOGIC INFORMATION SYSTEMS

Committee:

David Maidment, Supervisor

Timothy Bonner

Kenneth Dunton

Robert Gilbert

Ben Hodges

Daene McKinney

THE LONG TAIL OF HYDROINFORMATICS: IMPLEMENTING BIOLOGICAL AND OCEANOGRAPHIC INFORMATION IN HYDROLOGIC INFORMATION SYSTEMS

by

Eric Scott Hersh, B.S.C.E.; M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2012

Acknowledgements

I thank my advisor, David Maidment, for his vision, support, enthusiasm and guidance. I thank my other committee members and acknowledge their valuable insight, input, and support. I thank the Texas Commission on Environmental Quality and the Bureau of Ocean Energy Management for supporting this research. I thank Tim Whiteaker and Center for Research in Water Resources colleagues and staff. I thank the captain and crew of the R/V Alpha Helix and R/V Moana Wave for safe and successful field seasons. Lastly, I acknowledge the many valuable stakeholders, agency reviewers, workgroup members, and project teams who contributed to this research.

The research presented here is highly collaborative in nature. This research represents original thinking and writing on behalf of the author, but the following individuals are thanked for their specific assistance.

- Lisa Barden (Cockrell School IT Group) – digital library evaluation
- Bryan Enslein (CRWR) – EFIS data services
- Scott Hammock (CRWR) – EFIS data services
- Wendy Harrison (CRWR) – HydroPortal and EFIS data services
- Rick Hooper (CUAHSI) – ontology revisions
- Kate Marney (CRWR) – digital libraries
- Robyn Rosenberg (UT-Austin Engineering Library) – digital libraries

- Amy Rushing (UT-Austin Library) – digital libraries
- Harish Sangireddy (CRWR) – COMIDA CAB website and data management, Texas seagrass website
- James Seppi (CRWR) – data themes, EFIS website, map, and data services
- Clark Siler (CRWR) – Calculator for Low Flows
- Ryan Steans (Texas Digital Library) – digital libraries
- Tim Whiteaker (CRWR) – data model development, COMIDA CAB data management, EFIS data services
- Fengyan Yang (CRWR) – COMIDA CAB data management

The Long Tail of Hydroinformatics: Implementing Biological and Oceanographic Information in Hydrologic Information Systems

Eric Scott Hersh, Ph.D.

The University of Texas at Austin, 2012

Supervisor: David Maidment

Hydrologic Information Systems (HIS) have emerged as a means to organize, share, and synthesize water data. This work extends current HIS capabilities by providing additional capacity and flexibility for marine physical and chemical observations data and for freshwater and marine biological observations data. These goals are accomplished in two broad and disparate case studies – an HIS implementation for the oceanographic domain as applied to the offshore environment of the Chukchi Sea, a region of the Alaskan Arctic, and a separate HIS implementation for the aquatic biology and environmental flows domains as applied to Texas rivers. These case studies led to the development of a new four-dimensional data cube to accommodate biological observations data with axes of space, time, species, and trait, a new data model for biological observations, an expanded ontology and data dictionary for biological taxa and traits, and an expanded chain-of-custody approach for improved data source tracking. A large number of small studies across a wide range of disciplines comprise the "Long

Tail" of science. This work builds upon the successes of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) by applying HIS technologies to two new Long Tail disciplines: aquatic biology and oceanography. In this regard this research improves our understanding of how to deal with collections of biological data stored alongside sensor-based physical data. Based on the results of these case studies, a common framework for water information management for terrestrial and marine systems has emerged which consists of Hydrologic Information Systems for observations data, Geographic Information Systems for geographic data, and Digital Libraries for documents and other digital assets. It is envisioned that the next generation of HIS will be comprised of these three components and will thus actually be a Water Information System of Systems.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. BACKGROUND

The digital era has brought about a deluge of water information. Today's satellites, flux towers, aircraft, instruments, and ships are capable of monitoring the water environment with unprecedented spatial and temporal density, and today's high-performance computers are capable of processing tremendous numbers of operations for complex modeling and visualization. A growing world population coupled with an increasing global standard of living results in a pattern of increasing demand on the world's finite freshwater resources. Operating under the model that better information leads to better science and better decisions, Hydrologic Information Systems are emerging to organize, share, and synthesize this wealth of water information, but much of this information is held in independent databases that are unconnected.

If hydrologic science is taken to be "the science that treats the waters of the Earth, their occurrence, circulation and distribution, their chemical and physical properties, and their reaction with their environment, including their relation to living things" (Maidment 1993), then Hydrologic Information Systems are entities which store and transmit information that describes the properties of water and its motion through the earth system. In the sense used here, hydrology includes all aspects of the global water cycle, both terrestrial and marine, both above and below ground, and also incorporates the impact of water on living systems. This is a broader definition of hydrologic science than

that usually employed since it incorporates all the waters of the earth including oceanic waters and is not limited to just the waters directly associated with the land system.

In this new "sensor era" of the Information Age, a small number of scientific fields generate the vast majority of new data – particularly high energy physics, astronomy, climate modeling, and genomics. Research projects in these "Big Science" fields have terabyte- to petabyte-scale, relatively homogenous datasets and specific resources to deal with them – often million to billion dollar funding and hundreds to thousands of parallel processors. However, a great deal of useful research takes place outside of Big Science in a far larger number of smaller studies across a wide range of disciplines –the "Long Tail" of science (Figure 1). The term 'Long Tail' is derived from the probability distribution of data volume generated across various scientific disciplines, assumed to follow a power law distribution.

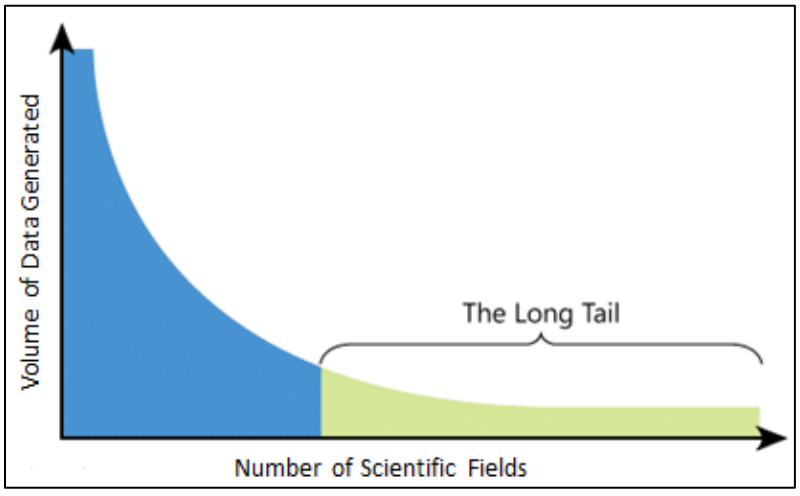


Figure 1. Few scientific fields – Big Science – produce the largest volume of data (blue area) and many fields – the Long Tail – each produce a smaller data volume (green area) (adapted from Chong and Carraro 2006).

The same data distribution has emerged in the hydrologic sciences. A small number of large-scale sensor networks and earth-observing systems produce sizable data streams describing the conditions of the planet's atmosphere, hydrosphere and cryosphere. These projects tend to originate in the national research labs and federal agencies of developed nations and typically have the financial, technical, and human resources for adept data management.

But a far larger number of hydrologic studies take place at the universities, labs, and other similar water research organizations of the world. These entities often collect "wet" data – samples from the field or the lab – which are heterogeneous and from ad hoc studies largely driven by funding availability and specific project needs. Compared to Big Science data on a per-data value basis, data in the Long Tail is often more expensive to collect and more difficult to curate. Yet these data have much potential value: as reference data, for validation, for transparency, for reuse, for aggregation and synthesis; the collective value of Long Tail data is enormous.

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) has advanced the study of hydrologic science in the United States over the past decade. The CUAHSI Hydrologic Information System (HIS) project has been successful in designing and deploying a national academic prototype Hydrologic Information System. CUAHSI HIS has greatly improved data access and data synthesis for the physical and chemical characterization of hydrologic systems ranging in scope from broad-scale (such as the United States Geological Survey streamflow monitoring network) to site-scale (such as the Critical Zone Observatory sites). While successful, the

CUAHSI HIS efforts to-date have largely excluded biological observations of the water environment and have largely been constrained to terrestrial hydrologic systems.

1.2. RESEARCH GOAL

Hydrologic Information Systems have emerged to address the data management needs of the Big Science elements of hydroinformatics, but very little work has been done to address the Long Tail data community within the hydrologic sciences. ***My work seeks to extend the current capabilities of Hydrologic Information Systems in order to provide additional capacity and flexibility for marine physical and chemical observations data and for freshwater and marine biological observations data.***

Consequently, one major goal of my work is to add biology to what has largely been a physical and chemical discussion and thus to make a step toward the integration of physical, chemical, and biological information for the water environment in a consistent and accessible manner in one system in one place. Furthermore, this work seeks not only to *define* how such information systems should look, but to actually *implement* prototype freshwater and marine systems, thus adding an oceanographic element to Hydrologic Information Systems which have thus far focused only on terrestrial elements.

1.3. PROBLEM STATEMENT

The nature of the problem being addressed by the research presented herein is the difficulty of accessing and synthesizing biological and oceanographic data for the water environment, and the solution put forth is improved organizing systems and tools. The challenges of water data management were recently summarized in the keynote address of the American Water Resources Association Spring Specialty Conference on GIS:

> "Water observations data are stored in many distributed tabular databases, each having its own output data format. Commonly measured variables such as streamflow or dissolved oxygen are labeled differently from one organization to another. The tabular databases are independently managed, not spatially enabled, and have no over-arching community or sponsor. In large water agencies, it can occur that data for different geographic regions are managed independently from one district or field office to another. The implication of this vast heterogeneity of water data systems is that data access and integration is laborious, so much so that perhaps 80% of an analyst's time is spent acquiring and processing data into useful forms before analysis can be carried out. As a result, water resources information is not leveraged as much as it could be, and big problems are not addressed effectively." (Dangermond and Maidment 2010).

These challenges are evident in Texas, where the Texas Commission on Environmental Quality (TCEQ) is tasked with developing instream flow recommendations. However, no comprehensive database of information is available for review and there is no systematic method for identifying or classifying Texas streams in order to determine the applicability of existing methods. Moreover, Texas Senate Bill 3 (2007) tasks stakeholders and regulators with determining and reviewing environmental flow needs, yet no repository of relevant data exists that could be shared with these stakeholders as they embark on the tasks of reviewing existing data and developing technical recommendations.

These same challenges hold true in an interdisciplinary Arctic Ocean project, and are even more evident when considering biological data for the water environment. Typically, aquatic biology studies involve more data complexity, data stored by individual investigators in their own way, and less effort expended on providing the necessary organizing systems and tools across projects. For every federal agency with redundant servers and a formal relational database structure, there are myriad researchers with "Dark Data" – folders of Excel files on their personal computer desktop, or, even worse, stacks of field data sheets tucked away in a file cabinet. In other words, everyone has data, few have databases.

A session was convened at the 2011 American Geophysical Union (AGU) Fall Meeting discussing "data scientists," an emerging role which defines those who can effectively communicate with both domain specialists (such as scientists and ecologists) and data managers (such as database experts and Information Technology (IT)

specialists). Data scientists are the practitioners who implement such Hydrologic Information Systems. This dissertation is written from the perspective of a data scientist.

1.4. SCOPE

The research proposed herein includes two case studies: the Chukchi Sea Offshore Monitoring in Drilling Area – Chemical and Benthos (COMIDA-CAB) project and the Texas environmental flows program. These case studies led to the development of a new four-dimensional data cube to accommodate biological observations data with axes of space, time, species, and trait. These case studies and the supporting research were conducted and are presented in such a way that scope increases successively and complexity is added successively – a "V-shaped" exposition from the Arctic to Texas.

In the first case study, the Chukchi Sea Offshore Monitoring in Drilling Area: Chemical and Benthos project (COMIDA CAB) is a robust, comprehensive effort to characterize the lease area biota and chemistry and to conduct a baseline assessment of the continental shelf ecosystem off the northwest coast of Alaska. A particular focus in this study is on ship-based physical, chemical, and biological sampling of the benthos and on the development of a workable food web model. As can be expected from such a multi-disciplinary effort, data management is an important and potentially challenging task and it is especially critical that a project-scale rather than investigator-specific database is developed. The reason for having the COMIDA CAB survey is to characterize the water and benthic conditions in the Chukchi Sea prior to drilling for oil

there. As such, it is likely that, decades into the future, comparisons will need to be made between historic and current Chukchi Sea conditions to assess the environmental impact of oil and gas exploration and production. To meet the project's data management needs, a data manager was ship-board to provide real-time, field-based data services and Geographic Information System (GIS) support. The author of this study established this role in the COMIDA CAB project.

In the second case study, stakeholders and regulators across Texas are in the midst of a legislatively-driven process to determine the environmental flow needs of the bays, basins, and rivers of the state. As is common elsewhere, the environmental flow program in Texas includes analyses of hydrology and hydraulics, geomorphology and physical processes, water quality, biology, and the connectivity between and among these four primary disciplines. The integration of sometimes disparate findings from these disciplines stands to be one of the most challenging and most important steps in developing instream flow recommendations. The Environmental Flows Information System for Texas created in this research seeks to provide improved data access and integration to aid stakeholder committees, expert science teams, and the Texas Commission on Environmental Quality in their collective efforts to determine statewide environmental flow needs.

To address the challenges posed by these two case studies and to help fill in these gaps, this research puts forth the concept of a next-generation Water Information System of Systems comprised of three components:

1. Geographic Information Systems (GIS) for geographic data,

2. Hydrologic Information Systems (HIS) for observations data, and
3. Digital Libraries for digital assets (documents, images, videos).

This research is novel in its integration of physical, chemical, and biological data describing the freshwater and marine ecosystem. As such, it seeks to help advance the maturity of the field of hydroinformatics.

1.5. RESEARCH QUESTIONS

In light of the above challenges, my research seeks to answer the following questions:

1. ***How can existing Hydrologic Information Systems which focus largely on physical and chemical data be made more robust to accommodate biological data?***

This research questions is addressed via an examination of the issues associated with biological data integration, the conceptualization of a data model for biological information, an elaboration of use cases and scenarios, and improvements and expansions to the information model currently in use for Hydrologic Information Systems.

2. ***How can existing Hydrologic Information Systems which focus largely on observations of the terrestrial water environment be made more robust to accommodate oceanographic data?***

This research questions is addressed via an analysis of the nature of oceanographic observations data and a comparison with the nature of terrestrial aquatic data, discussion of observing the ocean environment, organizing and storing oceans data, and communicating the results.

3. ***Is there a common framework for water information management for terrestrial and marine systems?***

This research questions is addressed via a detailed literature and technology review of existing tools and systems, an investigation and assessment of digital library technologies, and the introduction of a more robust 'system of systems' for water information which can accommodate geographic data, observations data, documents, and other digital assets (as opposed to existing systems which can only accommodate point observations data).

These three research questions are addressed in two case studies – an HIS implementation for the oceanographic domain as applied to the Chukchi Sea, located in the Arctic Ocean off the northwest coast of Alaska, and a separate HIS implementation for the aquatic biology and environmental flows domain as applied to Texas rivers.

1.6. WATER RESOURCE CHALLENGES

There exist multiple challenges in water resources today. Aquatic habitats are susceptible to habitat degradation and many aquatic species are globally imperiled. Imperiled species are those which are classified as threatened, endangered, or vulnerable. Habitat loss is listed as the primary threat in 85% of the species listed on the International Union for Conservation of Nature (IUCN) Red List (IUCN 2011). The Red List includes 17,000 imperiled species worldwide; of the 32,000 estimated fish species, 9,400 were evaluated and over 2,000 (21%) were deemed to be threatened (IUCN 2011). The picture is even bleaker in the United States– a 2008 study performed by the American Fisheries Society and the USGS found 39% of North American fish species to be imperiled (Jelks et al. 2008). The challenge here is one of habitat protection and restoration.

Further compounding other challenges faced by aquatic organisms, a changing climate is placing additional strain on numerous organisms and ecosystems the world over and threatens to materially affect water resources planning and management. The average surface temperature has risen 0.8 ºC in the last 100 years and is expected to rise an additional 1.5 to 6.1 ºC by 2100 based on varying emissions scenarios presented in the 2007 Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (Pachauri and Reisinger 2007). Recent discussion over 'the end of stationary' has many water managers concerned over the extent and timing of impacts to glaciers, precipitation regimes, extreme weather, and snow pack (Milly et al. 2008). The engineering challenges here are myriad – better forecasting and assessment, better adaptation

strategies, better response preparedness, better long-term preservation, and perhaps even more extreme responses such as geoengineering.

These big problems require big solutions and comprehensive global databases, but they also require detailed, specific solutions. The research discussed here seeks to add detail to the regional-to-global coverage of aquatic biology information in freshwater and marine environments.

1.7. DISSERTATION OUTLINE

This dissertation is organized into six chapters. Chapters two through five consist of four related papers which describe the research completed in response to the questions posed:

- Chapter 2 – Hydrologic Information Systems of the Past, Present, and Future
- Chapter 3 – Extending Existing Hydrologic Information Systems to Accommodate Biological Information
- Chapter 4 – Managing Arctic Marine Observations Data
- Chapter 5 – Managing Environmental Flows Information for Texas

Chapter 6 provides concluding remarks, discussion of how the research questions were addressed, and the anticipated contributions to science and engineering offered by this work. A glossary of acronyms is provided immediately following the conclusions.

The argument discussed here can be thought of as a V-shaped exposition (Figure 2). In the second chapter, a survey of past, present, and future of the field of

hydroinformatics is presented. The third chapter lays out the particular challenges associated with managing biological observations of the water environment and introduces something new called "the 4-D data cube." The fourth chapter addresses these biological data challenges for a relatively narrow, academic case study in the Arctic Ocean and brings hydroinformatics to the oceanographic realm. The fifth chapter branches out further to a case study which considers observations data for aquatic biology alongside other types of information and which serves a wider audience of stakeholders and practitioners in Texas in the field of environmental flows, plus a vision for the future of Hydrologic Information Systems is presented.

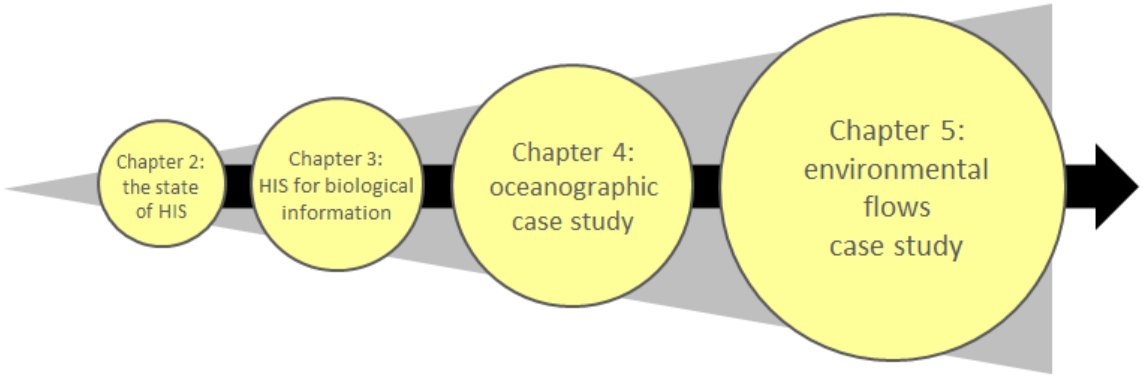


Figure 2. The dissertation narrative here can be depicted as a V-shaped exposition, where each successive chapter builds upon what was learned previously.

Chapter 2: Hydrologic Information Systems of the Past, Present, and Future

2.1. Existing Efforts

The field of hydroinformatics is relatively nascent. A thorough literature review was performed as part of this research encompassing dozens of hydrologic information systems, tools, projects, and efforts; *not one of them existed just ten years ago*. The general field for the research discussed herein is hydroinformatics: the science of information, the practice of information processing, and the engineering of information systems applied to water. Alternatively, hydroinformatics can be thought of as "the study of the flow of information related to the flow of water (and the entire water environment, in general)." (UNESCO-IHE 2010) Some of the work discussed herein lies at the interface of hydroinformatics and ecoinformatics; to understand the latter, simply replace "water" with "life." Note that some scientific communities, especially many in the European Union, use the term 'hydroinformatics' to encompass both the information science concepts discussed here as well as the field of computational fluid mechanics. The *Journal of Hydroinformatics* has been published since 1999, initially focusing on fluid mechanics aspects, but more recently broadening to incorporate research addressing both usages of the term. What follows is a broad survey of the state of hydroinformatics.

2.1.1 Cyberinfrastructure

"The term infrastructure has been used since the 1920s to refer collectively to the roads, power grids, telephone systems, bridges, rail lines, and similar public works that are required for an industrial economy to function. Although good infrastructure is often taken for granted and noticed only when it stops functioning, it is among the most complex and expensive thing that a society creates. The newer term cyberinfrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy." (Atkins et al. 2003)

If cyberinfrastructure is analogous to transportation infrastructure, carrying data instead of cargo, service-oriented architecture (SOA) is the Eisenhower Interstate Highway System plan. SOA is a design model based on services that communicate via a shared protocol. (Erl 2004, Erl 2005) It's variously a paradigm, perspective, or concept applied to large, distributed information systems where resources on a network are made available as independent services, decoupled from operating systems and platforms. (Josuttis 2007)

Web services "provide the ability to pass messages between computers over the Internet, therefore allowing geographically distributed computers to more easily share data and computing resources." (Goodall et al. 2008) Web services make use of standardized protocols to announce their capabilities and content (the Web Services Description Language, or WSDL), to communicate their content (via Simple Object Access Protocol, SOAP, or, increasingly, via Representational State Transfer, REST),

and to publicize their existence (via Universal Description, Discovery, and Integration, UDDI). Simply put, data access through REST is explicitly communicated through a Uniform Resource Locator (URL) while data access through SOAP is communicated via a set of objects contained within a WSDL. (Curbera 2002, Kumar et al. 2006) The communication of web services via SOAP and REST is now overwhelmingly accomplished using the Extensible Markup Language (XML) – the lingua franca for data encoding (Curbera 2002, Kumar et al. 2006). XML can be thought of as a generalization of the HyperText Markup Language (HTML). *Web pages* deliver text and images over the internet encoded as HTML, whereas *web services* deliver data over the internet encoded as XML.

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. Hydrologic Information System (CUAHSI HIS) team developed WaterML as a customized XML specifically for describing the data and metadata for physical and chemical water observations collected at point locations. CUAHSI was involved in the Open Geospatial Consortium/World Meteorological Organization (OGC/WMO) Hydrology Domain Working Group which designed an updated version of WaterML compliant with the international information standards set forth by those two bodies (Maidment 2009).

As a result of that Working Group's efforts, OGC adopted WaterML 2.0 part 1: Time Series Encoding Standard as an official OGC standard. (OGC 2012b) WaterML has an associated set of web services called WaterOneFlow which supports four client requests: (1) GetSites, for a list of sampling sites in a particular network; (2) GetSiteInfo,

for detailed site metadata, variables list, period of record, and value count for each variable; (3) GetVariableInfo, for variable metadata; and (4) GetValues, for a time series of data values at a given site for a given time period (Zaslavsky et al 2007). The United States Geological Survey (USGS) now publishes both daily streamflow data and unit values data (real-time information) from the National Water Information System (NWIS) in WaterML. (Maidment 2009)

No discussion of data would be complete without a discussion of metadata – data about data. Stored in a standardized format and transmitted via XML and WaterML, metadata describes the basic characteristics of data or information: the who, what, where, when, why and how (FGDC 2009). Although multiple standards exist, metadata for Geographic Information Systems in the United States are often published using the Federal Geospatial Data Committee (FGDC) Content Standard for Digital Geospatial Metadata or the International Organization for Standardization (ISO) 19115 Metadata Standard. (Kumar er al. 2006, FGDC 2009, ISO 2009) Examples of geospatial metadata include geographic extent, projection, and scale. Similarly, metadata for Digital Library contents are often published using the Dublin Core Metadata Initiative (DCMI) standards. (DCMI 2009) Examples of library metadata include title, abstract, publisher, and publication year.

Thus far, metadata for Hydrologic Information Systems are published within the CUAHSI Observations Data Model (ODM) structure, a customized metadata format. (Horsburgh et al. 2008) Important to note is the distinction between *storing* and *communicating* data; in CUAHSI's case, the ODM is used as a storage repository data

and metadata whereas WaterML is used as the transmission language to communicate the same data.

2.1.2 Geographic Information Systems

A Geographic Information System "integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically reference information." (ESRI 2009) Geographic data are typically static in time, complex in space, and are organized in standardized formats such as geodatabases. Geographic Information Systems were first conceptualized in the 1960s and now sustain a mature commercial market (Foresman 1998).

ESRI is the leading global provider of traditional GIS software and services. Founded in 1969, ESRI is, by many accounts, the research and development leader in the geospatial information market and the dominant player in the 'traditional GIS space' of desktop and enterprise software. It is interesting to note, however, that ESRI GIS products represent only approximately 40% of the GIS market share (Arc Advisory Group 2010). Furthermore, the advent of web mapping services such as Google Maps, Google Earth, and Microsoft Bing Maps has rapidly and dramatically reshaped the GIS market, and this new online GIS space dwarfs the traditional desktop GIS space. For example, Google estimates that there are over *500 million* active users of its mobile and web mapping applications (Siegler 2011).

The most common building block for geographic data today is the proprietary geodatabase from ESRI. Introduced in 1999 as part of the ArcGIS 8.0 release, a geodatabase is a collection of geographic elements stored within a relational database structure or in a file database structure. A geodatabase is comprised of: (1) feature datasets, collections of feature classes of vector-based geographic data with the topology and network objects supporting them; (2) tables of attributes; (3) relationships linking the tables and feature classes; (4) raster data for continuous geographic phenomena; and (5) metadata (Arctur and Zeiler 2004).

Data models provide the underlying structure to both Geographic Information Systems and Hydrologic Information Systems. Data models are a formal method of describing the behavior of real-world entities, "sets of concepts describing a simplification of reality expressed in database structures such as tables and relationships, and they provide standardized frameworks for users to store information and serve as the basis for applications." (Arctur and Zeiler 2004) Geographic data models are a special case of data model where spatial database structures are used and stored in a spatial database to describe geospatial phenomena using Geographic Information Systems. Put more simply, data models define objects of interest and identify relationships and geographic data models do this in a spatial context. In a GIS, geography is dominant and variable and time are subordinate; in an HIS, variable is dominant and geography and time are subordinate. That is to say, geography is the central focus of a GIS data model and the observation itself is the central focus of an HIS data model.

ESRI supports and maintains 34 data models in fields ranging from agriculture to defense to petroleum (ESRI 2010b). Arc Hydro is the data model for surface water resources, combining geospatial and temporal data within an ESRI geodatabase schema in order to support hydrologic analysis and modeling (Maidment 2002). Arc Hydro is built on a thematic framework of river and stream network, terrain elevation data, and watershed boundaries and includes a basic treatment of physical water observations data via a coupled point feature class and time series representation. Arc Hydro has been adopted, applied, and extended by a diverse user community and has recently been spawned a corresponding data model, Arc Hydro Groundwater, which extends the Arc Hydro surface water framework by introducing a representation of multi-dimensional ground water data, including geologic stratigraphy, hydrostratigraphy, aquifer maps, borehole data, and simulation model support (Strassberg et al 2007, ESRI 2010b, Strassberg et al. 2011).

A related data model has been developed to support observations in the ocean realm, Arc Marine. Arc Marine includes similar representations of vector, raster, and time series data as Arc Hydro but adds additional support for limited three-dimensional geographic data from model mesh volumes and also the unique feature of storing observations data collected from along a moving track, such as a ship towing a sensor measuring conductivity and temperature at various water depths (Wright et al 2007, ESRI 2010b). Streamflow data and other similar surface water observations are made at fixed point locations, as are many marine observations made from buoys, Acoustic Doppler Current Profilers (ADCPs), hydrophones and tidal gauges. However, it is not uncommon

for marine data to come from a mobile sampling platform such as a ship, drifter, autonomous underwater vehicle, or even a tagged animal. Arc Marine's schema has the capacity to store the observations themselves as well as the track and its attributes (Wright et al 2007).

2.1.3 Hydrologic Information Systems

A Hydrologic Information System is "a services-oriented architecture for water information" consisting of a repository of hydrologic time series data (HIS Server), a national water metadata catalog (HIS Central), and a desktop appliance for hydrologic data access (Hydro Desktop). (Maidment 2009) These three elements are linked via web services, "automated functions that enable one computer to make appropriate requests of another computer and receive responses through the internet." (Maidment 2009) Water observations data are typically dynamic in time (time series), simple in space (sampling and gaging points), lacking in standardized formats, and potentially stored in relational databases. Hydrologic Information Systems were conceived in the early 2000s by the National Science Foundation-supported Consortium of Universities for the Advancement of Hydrologic Science, Inc. and are a new and evolving concept; HIS development has thus far been accomplished by the CUAHSI university partnership with some business partners. (Maidment 2009)

The CUAHSI Hydrologic Information System is built around a normalized data storage schema called the Observations Data Model (ODM). (Horsburgh et al. 2008)

The ODM provides a consistent relational database format for storing point observations data and their supporting metadata in a manner which exposes each single measurement as a unique record with associated descriptors as to location and time of measurement and the method used, and which addresses many of the syntactic and semantic differences between heterogeneous data sets (Horsburgh et al 2008). The ODM logical data model features the data value itself in a central role with the supporting metadata attached to it via associated metadata tables – a 'star' schema. The central DataValues table includes a number of Foreign Keys which link to Primary Keys in other tables: site location, offset, units, variable, source, sampling method, lab method, data qualifiers, and quality control level, along with a compiled series catalog to facilitate indexing and searching (Figure 3).

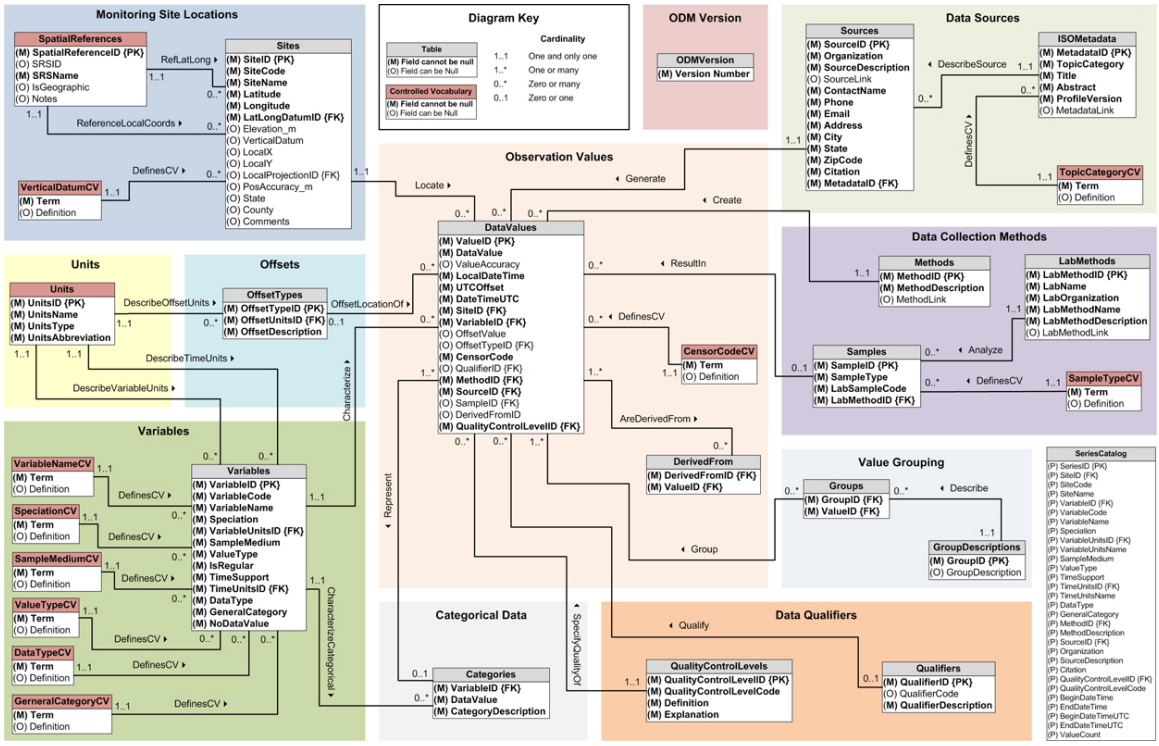


Figure 3. CUAHSI Observations Data Model schema (Horsburgh et al. 2008)

With support from the NSF Earth Science Division, the 125 member universities of CUAHSI developed a national academic prototype Hydrologic Information System. The national CUAHSI HIS has been adapted and implemented at a state-level in Texas – the Texas Hydrologic Information System. By using traditional and hybrid web services, over 23 million observations have been catalogued, encompassing over 7,000 variables from nearly 16,000 sites and 15 state-specific data providers (Whiteaker et al 2010). The Texas HIS presents the concept of thematic data organization, a synthesis across data providers via a region- or discipline-specific grouping. For example, four independent state agencies each collect salinity data along the Texas coast, and the aggregation of

these data likely provides additional value to a researcher. The prototype Texas HIS has been supported by the Texas Water Development Board and that agency is currently in the process of transitioning to a permanent production system (Whiteaker et al 2010).

Just as the US National Science Foundation is supporting the development of the CUAHSI HIS, the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) Land and Water is supporting the parallel development of the Australian Water Resources Information System (AWRIS). The two entities have largely similar goals and have enjoyed a partnership and cooperation, particularly in the development of WaterML 2.0 (CSIRO 2009). The US Environmental Protection Agency (US EPA) has an internal data management system somewhat akin to the Hydrologic Information Systems of CUAHSI and CSIRO. The Water Quality Exchange schema, or WQX, is a data storage and communication format designed to facilitate the aggregation of regulatory water quality data from states and tribes into a Central Data Exchange, then to a National STORET Data Warehouse (STORET is the EPA's legacy data STOrage and RETrieval system), then to be disseminated via web services and consumed by analysis and mapping applications (US EPA 2010).

A suite of tools and systems has emerged to manage the variety of information types present in the water environment, but significant gaps remain (Table 1). With respect to geography, the pioneering Arc Hydro data model for surface water (Maidment 2002) helped to shape the Arc Marine data model for oceanography (Wright et al. 2007) and spun-off the Arc Hydro Groundwater data model for hydrogeology (Strassberg et al. 2011). With respect to relational databases for physical and chemical observations, the

CUAHSI Observations Data Model (Horsburgh et al. 2008) was without specific counterpart for marine systems prior to the work presented herein. Similarly, relational databases for biological observations did not exist prior to this work.

Table 1. Prominent existing systems for freshwater and marine data management.

Data Type	Freshwater	Marine
Geographic	Arc Hydro (Maidment 2002) & Arc Hydro Groundwater (Strassberg et al. 2011)	Arc Marine (Wright et al. 2007)
Physical and chemical observations	CUAHSI Observations Data Model (Horsburgh et al. 2008)	--
Biological observations	--	--

2.1.4 Digital Libraries

A Digital Library is a collection of digital materials (as opposed to print, microform, or other physical media) accessible via computer. (Greenstein and Thorin 2002) Digital libraries, also known as digital repositories, provide for large-scale, stable, managed long-term storage of digital material in any format and are designed to capture, describe, distribute and preserve these materials. (Kainerstorfer and Perkins 2009) Digital content may include technical reports, articles, books, maps, tables, photographs, images, videos: any material which is either born-digital or digitized. Digital libraries

were conceptualized in the mid-1990s, sustain a modest commercial and open-source presence, and are the focus of more widespread experimentation and development. (Greenstein and Thorin 2002)

DSpace (http://www.dspace.org/) is a digital repository system developed by the Massachusetts Institute of Technology (MIT) Libraries and Hewlett-Packard Labs that captures, stores, indexes, preserves, and redistributes an organization's research data and is the repository system used by The University of Texas at Austin. Digital repositories such as DSpace allow organizations to organize and store a variety of data formats in an accessible and persistent manner. DSpace accepts content such as articles, technical reports, working papers, conference papers, theses, datasets, images, audio and video files, and reformatted digital library collections. DSpace operates on a logical infrastructure, utilizing metadata for organization and retrieval. Data files, also called bitstreams, are organized together into related sets. Each data file has a technical format and other technical information (DSpace 2008).

2.1.5 Digital Library Systems Review and Evaluation

As part of a demonstration Digital Library project for the river basin and bay system consisting of the Trinity and San Jacinto Rivers and Galveston Bay, a number of existing digital repository systems were reviewed and evaluated. These included: Knowledge Tree, Brazos River Instream Flows Study Database, Xythos Server Products, and Inmagic Presto (in addition to DSpace, described above).

Knowledge Tree (http://www.knowledgetree.com/) is a commercial open source, web-based document management system that is currently licensed by the Aerospace Engineering Department at UT-Austin and is best suited for document workflow management. (Knowledge Tree 2008) The Brazos River Instream Flow Study Database is a Microsoft Access database developed by Espey Consultants, Inc. that is used to document project reports. This application features a Microsoft Access interface and a GIS component for spatial representation. (Espey 2005) Xythos Server Products (http://www.xythos.com/products/webfile_server.html) are three software suites used to accommodate an institutional repository. (Xythos 2008) Inmagic Presto (http://www.inmagic.com/products/research/presto.html) is a Web-based application for accessing, sharing, and managing research information that is partnered with WebFeat to provide federated search capability across external data sources. (Inmagic 2008; Hersh et al. 2008)

2.1.6 Managing Ecological Information

While CUAHSI is viewed by many to be the main standard-bearer for hydroinformatics in the United States, a number of cyberinfrastructure projects are underway in the realm of ecoinformatics with widely varying degrees of maturity, support, scope, and acceptance; this situation might be a result of the wide diversity of researchers and projects under the ecology banner and of the previously-discussed challenges presented by biological data management.

The National Ecological Observation Network (NEON) is deploying a sensor network and cyberinfrastructure to support ecological research (NEON 2010). The Long-Term Ecological Research Network (LTER) program consists of a series of research sites designed to facilitate comparison and synthesis across diverse habitats (LTER 2010). The National Center for Ecological Analysis and Synthesis (NCEAS) has as its mission to serve as a data center for the ecology and evolutionary biology communities and has created a flexible metadata standard for the description of ecological data, the Ecological Markup Language (EML) (Madin et al 2007). EML's allowed flexibility in format and content and its design goal to support data discovery differs fundamentally from the rigid specifications of WaterML, a format designed to support both data discovery and integration.

Similarly, the USGS National Biological Information Infrastructure (NBII) is working to provide national metadata standards unique to biological information (Ruggiero et al 2005). The non-profit group NatureServe is developing a network of natural heritage programs and conservation data centers in the Western Hemisphere for the application of ecoinformatics to conservation science and policy. In support of this effort, the Biotics4 Physical Data Model has been created to store ecological observations data and web services to communicate these data are being developed (NatureServe 2010). Finally, the Global Biodiversity Information Facility (GBIF) is an international effort to provide a cyberinfrastructure for information on the world's biodiversity data. GBIF largely focuses on cataloging the described species of the world and recorded observations of those species (Edwards 2000, GBIF 2010).

2.1.7 Managing Aquatic Biology Information

As discussed, there exist a diverse range of ecoinformatics efforts, some of which are relatively mature and some of which are very well-funded. In the specific realm of biological observations for the water environment, there are many fewer projects with a collectively lesser degree of maturity.

The most well-known and well-used such effort is FishBase, an online collection of fish observations from across the globe, boasting "31,500 species, 279,900 common names, 49,200 pictures, 43,800 references, 1750 collaborators, and over 33 million hits per month." (Froese and Pauly 2010). Started in 1995 as a CD-ROM, the goal of FishBase (as in, 'Fish Database') is to host and serve information on the biology, distribution, and taxonomy of the world's fishes (McCall and May 1995, Froese and Pauly 2010).

The North-Temperate Lakes Long Term Ecological Research site (NTL LTER) in Wisconsin has developed a database and an internal data model for storing observations of physical and chemical limnography plus some aquatic organism observations and collections, ranging from plankton to fish. The data model for fish includes: LakeID, Year, SampleDate, GearID (the equipment used to sample), SpName (species common name), SampleType, Depth, Rep (Replicate Number), Indiv (individual ID), Length, Weight, Sex, FishPart (the portion of the fish analyzed or sampled, such as stomach, scale, or otolith), and SpSeq (species sequence number). Similarly, the benthic macroinvertebrate data model includes: LakeID, Year, Site, Rep, Taxon_Code (taxonomic code), Description (taxonomic code description), Number_Indiv (number of

individuals collected), and Flag (data flag for number counted); embedded within the taxon code is an eight-tiered taxonomic classification, from phylum down to genus (NTL LTER 2008).

The ISEMP Aquatic Resources Metadata Framework was created by the consulting company Environmental Data Services of Portland, Oregon for the Integrated Status and Effectiveness Monitoring Program of the Northwest Fisheries Science Center. Essentially a customized Observations Data Model for fish observations in the US Pacific Northwest, the Framework includes fields for the fish (redd presence and condition, electrofishing details, fish attributes, size class, taxonomy, and genetics); habitat (physical parameters, survey station, transect, water quality, cover, large woody debris, riparian vegetation, alteration, and substrate); data collection event (weather, samplers, equipment, and data provenance); protocol (sampling methodology); site (diagram, map, survey, position, geographic setting, and hydrographic setting); statistical design; and project (Environmental Data Services 2008).

Two other such internal biological data management systems exist. The first is the Ecological Data Application System (EDAS) developed by Tetra Tech, Inc. EDAS is designed to facilitate data analysis for the multi-metric indices commonly calculated and used in the study of benthic macroinvertebrates, particularly in relation to environmental monitoring and assessment efforts. EDAS is built on a Microsoft Access-based relational database of benthics, water chemistry, and physical habitat and includes index calculations and an export to STORET feature (Tetra Tech 2000). The second is the Freshwater Biodata Information System (FBIS) of the New Zealand National Institute of

Water and Atmospheric Research (NIWA). FBIS is the database which resulted from a reorganization of decades of internal biological data collection for fish, invertebrates, submerged macrophytes, and bryophytes/algae/diatoms. FBIS supports some search functionality and an interactive map viewer (Robertson and de Winton 2004).

2.1.8 Managing Marine Observations Data

At another state of database maturity, we arrive at the marine biology community. Biological oceanographers have a number of large-scale, well-received data archives, notably the National Oceanographic Data Center (NODC) and National Center for Atmospheric Research Earth Observing Laboratory (NCAR EOL), but these archives store datasets; very few databases and/or data models exist for biological observations in the marine environment. Both the NODC and NCAR EOL archives include a wealth of researcher-submitted data for a wide range of physical, chemical, and biological oceanographic observations. However, data is welcomed in any native physical format or structure and is archived as such; no efforts toward synthesis or integration are evident (NODC 2010, NCAR EOL 2010).

Data from the Western Arctic Shelf-Basin Interactions (SBI) project of the National Science Foundation are an example of project data stored in the EOL archive; a brief investigation of the SBI data archive yields various data provided in txt, pdf, gif, and xls file formats with access via html and ftp; no standardization appears evident (SBI 2008).

There are some nascent efforts toward developing cyberinfrastructure for the ocean realm, however. One such effort is the Ocean Biogeographic Information System – Spatial Ecological Analysis of Marine Megavertebrate Animal Populations, or OBIS-SEAMAP. This project seeks to develop a geodatabase of sea turtle, marine mammal, and seabird global distribution and abundance data (Halpin et al 2006).

2.2 CURRENT EFFORTS

2.2.1 Current Efforts in Hydroinformatics

The CUAHSI HIS project grant ended on December 31, 2011, but research on Hydrologic Information Systems continues elsewhere, particularly in three new projects. First is EarthCube: "A collaboration between the U.S. National Science Foundation (NSF) and geo, atmosphere, ocean, computer, information, and social scientists. EarthCube aims to transform the conduct of research through the development of community-guided cyberinfrastructure to integrate information and data across the geosciences." (EarthCube 2012) Second is HydroShare, an NSF-sponsored cooperative effort between the Renaissance Computing Institute at UNC Chapel Hill (RENCI), Utah State University, and six other university partners to expand the work of the CUAHSI HIS project as it relates to the particular focus of data sharing (RENCI 2012). HydroShare is focused on providing an online community for hydrologists who want to work collaboratively sharing data and models (Figure 4).

Figure 4. A mock-up of the HydroShare community data sharing interface (RENCI 2012).

Third is CI-WATER, where an interdisciplinary team from Utah and Wyoming in seeking to "develop a better understanding of the interconnectivity of natural and human water resources systems – a critical environmental sustainability problem facing both Western states." (CI-WATER 2011)

2.2.2 Current Efforts in Spatial Data Infrastructure

In addition to the focused hydroinformatics efforts discussed above, a number of broader efforts are underway with the goal of organizing and advancing the global spatial data infrastructure. Primary among these global efforts is the Global Earth Observing System of Systems (GEOSS) (Figure 5). A 10-year effort initiated in 2005, GEOSS "seeks to connect the producers of environmental data and decision-support tools with the end users of these products, with the aim of enhancing the relevance of Earth observations to global issues. The result is to be a global public infrastructure that generates comprehensive, near-real-time environmental data, information and analyses for a wide range of users." (GEO 2012).

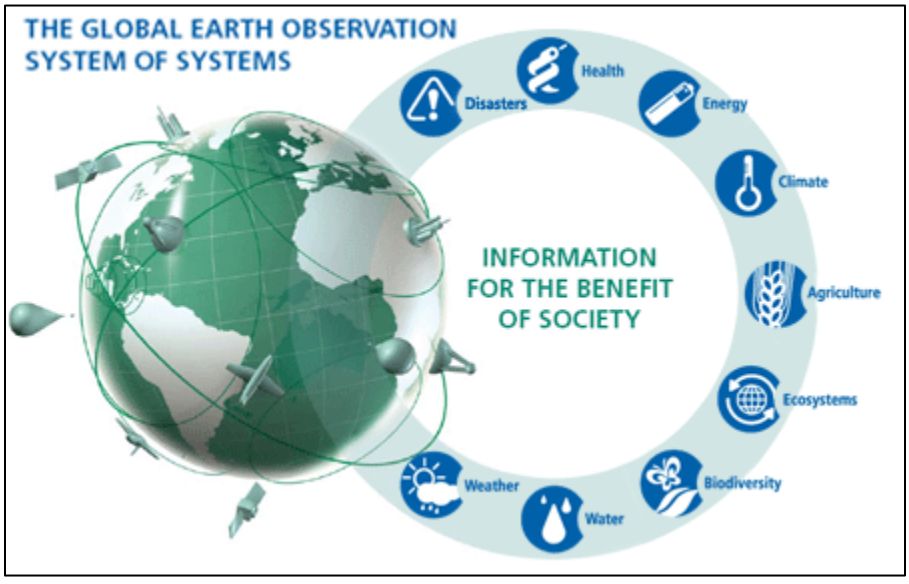


Figure 5. The Global Earth Observing System of Systems (GEO 2012).

GEOSS is being developed by the Group on Earth Observing (GEO), whose membership includes 75 nations and 51 partner organizations and which was founded as an outcome of the 2002 World Summit on Sustainable Development by the Group of Eight (G8) leading industrialized nations. The GEOSS "'system of systems' will proactively link together existing and planned observing systems around the world and support the development of new systems where gaps currently exist" and also seeks to provide a "GEO Portal" for online data access (GEO 2012). GEOSS seeks to serve nine "societal benefit areas:" disasters, health, energy, climate, agriculture, ecosystems, biodiversity, water, and weather via the linkage of existing and planned observing systems.

2.3 DATA-INFORMATION-KNOWLEDGE-WISDOM

A popular model in information theory is the Data-Information-Knowledge-Wisdom Pyramid, also variously known as the DIKW Hierarchy, and the Knowledge Hierarchy (Ackoff 1989) (Figure 6). This model holds that Data is the raw facts, Information gives meaning to Data, Knowledge is analyzing and synthesizing Information, and Wisdom is using Knowledge to establish and achieve goals (Baker 2007, Elias 2011). Under this model, value is added at each step of the hierarchy.

An interpretation of the DIKW Pyramid was developed specifically for water information – The Water Information Value Ladder (Vertessy 2010) (Figure 7). In this

interpretation, Data leads to Information which leads to Insight, each possessing increasing value.

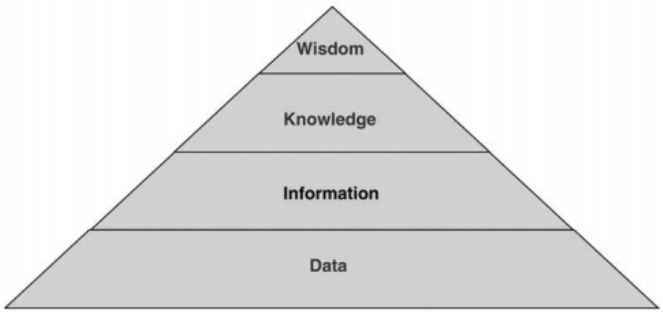


Figure 6. The Data-Information-Knowledge-Wisdom Pyramid (Rowley 2007).

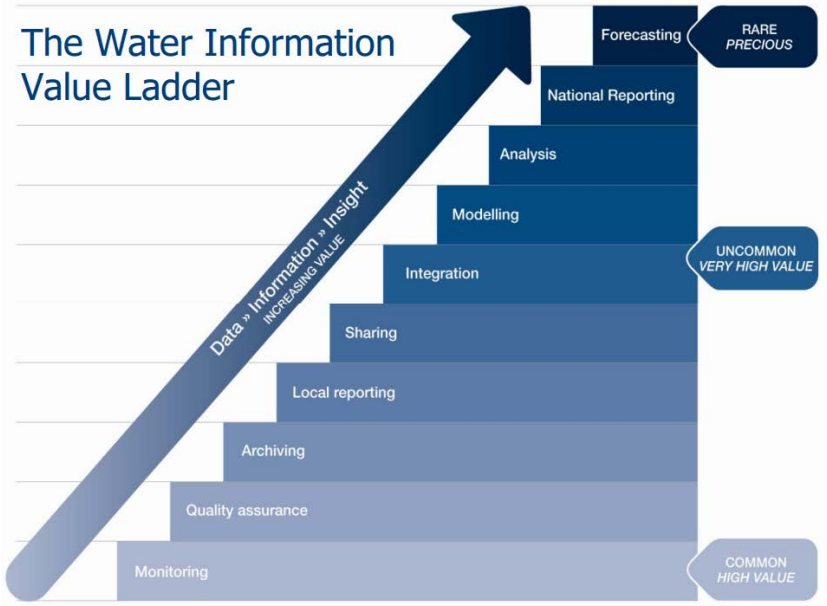


Figure 7. The Water Information Value Ladder (Vertessy 2010).

Here, the relative maturity of existing hydroinformatics tools and systems from varying domains and disciplines is presented in similar fashion (Figure 8). Is it hoped that, as with the DIKW Pyramid and the Water Information Value Ladder, additional value will be realized as each of these efforts develop and mature. This could be in the commercial sense, as dollars and cents, in the conservation sense, as increased awareness and protection of our limited resources, or in the research sense, as increased understanding of the processes and mechanism of the world around us.

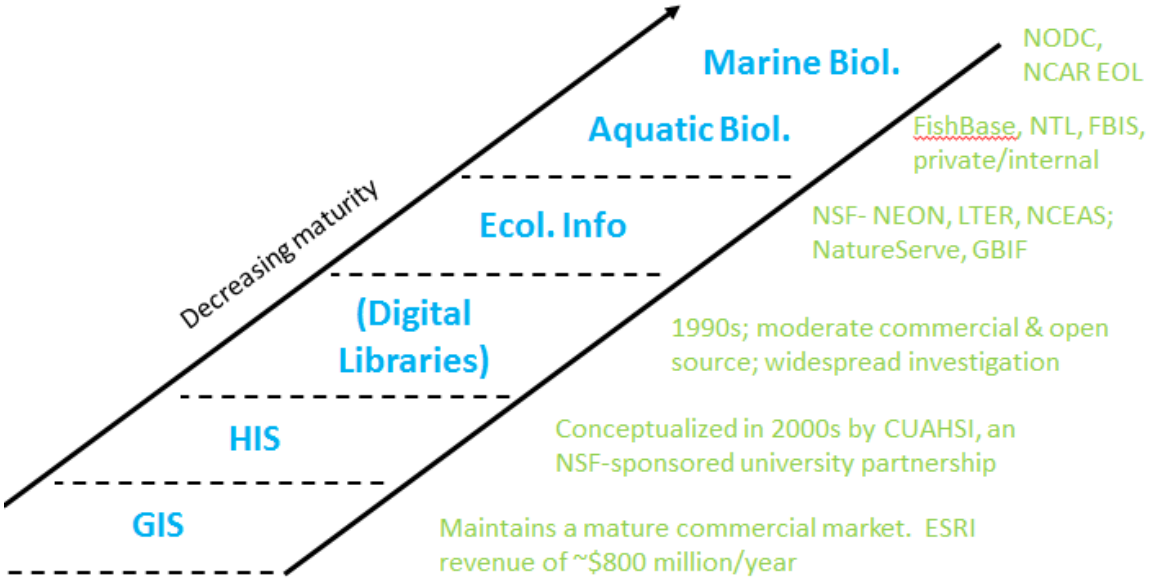


Figure 8. The hydroinformatics maturity ladder (adapted from Vertessy 2010).

2.4 KNOWLEDGE MANAGEMENT

The approach presented here is a step toward more complete "knowledge management," the sharing of not just data but of information and insight derived from those data (Alayi and Leidner 2001). A common viewpoint is that "data is raw numbers and facts, information is processed data, and knowledge is authenticated information," that which has been "actively processed in the mind of an individual through a process of reflection, enlightenment, or learning." (Alayi and Leidner 2001)

For the dual purposes of knowledge management and quality control, a "data chain" is envisioned which connects information in publications with the data sources from which it is drawn, thus linking information in a report back to the data sources upon with the figures and tables in the report were prepared (Figure 9). In this vision, all types of project information relevant to the analysis, reporting, and publication of results are accommodated.

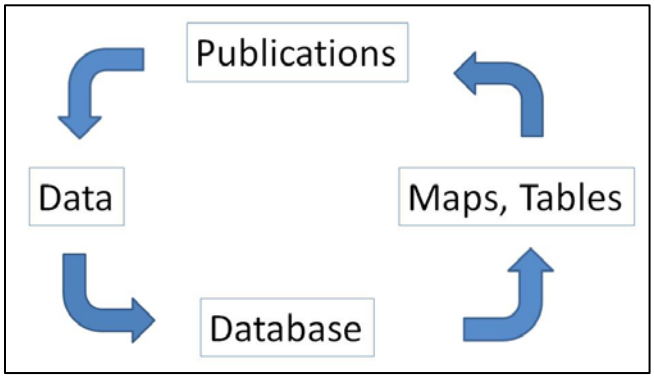


Figure 9. The "data chain" vision.

2.5 CONCLUSIONS

This chapter has presented an in-depth look at the suite of systems, tools, projects, and efforts in the relatively nascent field of hydroinformatics. Significant advances have been made in information management within each 'pillar' of data, databases, data themes, and digital assets via such technologies and Hydrologic Information Systems, Geographic Information Systems, and Digital Libraries. But very little work has been accomplished to-date in holistic data management – adequately organizing and storing information of differing types. More complete data integration of different data types represents a step toward more complete "knowledge management," where data is presented alongside the information and insight derived from those data. The following chapters will describe advances made toward that goal.

Chapter 3: Extending Existing Hydrologic Information Systems to Accommodate Biological Information

3.1. The Water Environment

The importance of water cannot be overstated. "Water is the most abundant substance on earth, the principal constituent of all living things, and a major force constantly shaping the surface of the earth." (Chow et al. 1988) As such, numerous disciplines have developed to study various aspects of the water environment. Hydrology is the study of water as a physical environment – the movement and distribution of water through the land and air. Aquatic biology and marine biology are the study of water as a living environment – the habitat and organisms that live in freshwater systems and marine systems, respectively. Finally, water resources is the study of water as it pertains to human need – providing and maintaining water for drinking, for agriculture, for industry.

Accordingly, there are distinct types of water data, each type with its own character. Physical data describe the movement of water and its properties. Chemical data describe the constituents moving with, in, and through the water. Biological data describe the organisms inhabiting the water environment. Subsets and hybrids of these broad domains exist; for example, bathymetric and geomorphic data describe the physical environment of the water but also the geochemistry and the habitat. As is often the case

for hydrology, the spatial and temporal extent of data required for meaningful analysis likely exceeds that which can reasonably by accomplished by individual project-specific data collection efforts. Data collection in a freshwater or terrestrial setting is a very different experience than in a marine or offshore setting. And data from a federal observation network is different than data from a community volunteer monitoring organization.

This chapter lays out the particular challenges associated with managing biological observations of the water environment. This is accomplished through an investigation into the characteristics of biological observations data and both conceptual and practical means of organizing these data. A novel conceptual approach to accommodating both organism taxonomy and traits called the 4-D data cube will be introduced, as will a new observations data model custom-designed for biological observations.

3.2. THE NATURE OF BIOLOGICAL INFORMATION

As a result of the typical complexity of biological information and the limitations of biological data collection, biological data is distinguished from physical and chemical data by a number of aspects. Much of the biological information for the freshwater and marine ecosystems is collected by techniques including: grab samples, electrofishing, net hauls, and population surveys – distinct sampling 'events.' Researchers visit a field site, deploy gear, collect a sample, and process that sample on-site or in a lab. These discrete

collections are in contrast to continuous time series collections where a sensor, gage, or other real-time or data logging instrument collects a large volume of data at regular intervals. Thus, the resources required to collect biological samples often render biological data more irregular in time and space and generally less voluminous than physical or chemical data for the water environment (Table 2).

Table 2. Data characterization and comparison.

Data Type	Water Resources	GIS	Aquatic Ecology
Temporal	dynamic (time series)	static	event-based (irregular)
Spatial	simple (points)	complex	complex (3-D)
Format	non-standardized	standardized	non-standardized & compound (data interplay)

Biological data is an important component of the "data ecosystem" and contains a specific kind and organization of information that attempts to capture the very considerable complexity of biological processes. There are hundreds of physical parameters, thousands of chemical constituents, and millions of biological species affected by water systems. Biology is necessarily more complex than physical and chemical characterization of water properties because it deals with living systems whose

species are interacting with one another and with the environment in which they are immersed.

Although the methodology is labor-intensive to get each data value when compared to the elaborate continuously-monitoring sensor networks in use today, biological collections provide additional information content based on the parameters measured during a sampling event. For example, a common riverine fisheries data collection effort might include sampling to determine the size and characteristics of multiple species of fish plus a suite of sampling to determine the water quality, flow conditions, and physical habitat which support that particular fish community.

The physical and chemical conditions are strong drivers of the diversity, size, and character of the fish community, as are the different types of fishes present (predators and prey, invasive species, tolerant and intolerant species, varying life stages, etc). Thus, the interplay of the observations made during a sampling event can provide considerable value, and another important distinction is drawn: the lateral (i.e., multiple variables measured as part of one sample) nature of biological data typically characterizes additional value of that data, whereas the longitudinal (i.e., one variable measured through time) nature of physical and chemical observations through time typically provides the information content in those realms.

In summary, biological data management is to a collections-based framework as physical data management is to a series-based framework; chemical data management is an intermediary, consisting of both sampled collections (the majority of effort today) and continuously monitored information.

3.3. TAXONOMIC CLASSIFICATION

There are 1.2 million distinct species currently catalogued on earth (Bisby et al. 2010). Estimates of the actual number of species on earth range from 3 to 100 million, and one popular estimate posits that there are actually as many as 8.7 million species globally, of which approximately 2.2 million species reside in the ocean (Mora et al. 2011). Assuming this estimate is relatively accurate, only 14% of terrestrial species and only 9% of marine species have been catalogued!

Modern taxonomic classification was introduced by Swedish scientist Carolus Linnaeus in 1735 in his book *Systema Naturae* and the current incarnation of Linnaean classification was set forth in the 10th edition of that book (1758). Modern taxonomists group organisms based on shared traits and categorize these groups hierarchically, with some variation within and among the defined level (Figure 10). Recent advances in DNA sequencing and genomics have led to a new super-tier of classification, the Domain (Woese et al. 1990). Most biologists now define three Domains:

1. Bacteria – prokaryotes, mostly unicellular, whose cells lack a nucleus or any other organelles;
2. Archaea – single-celled prokaryotic microorganisms with separate evolutionary history and distinct genetics from bacteria; and
3. Eukarya – organisms whose cells contain complex structures enclosed within membranes.

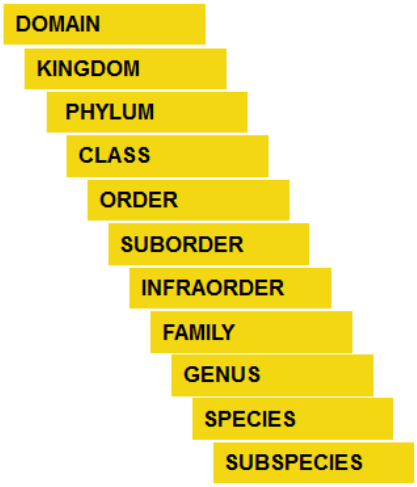


Figure 10. Example hierarchical taxonomic classification system.

A particular organism is unique defined by the binomial nomenclature of genus and species. For example, the largest animal on earth, the blue whale, is classified as *Balaenoptera musculus* (Table 3).

Table 3. Taxonomic classification for *Balaenoptera musculus* (blue whale).

Level	Classification	Description
Domain	*Eukarya*	complex cell structures enclosed within membranes
Kingdom	*Animalia*	animals
Phylum	*Chordata*	chordates (possessing a nerve cord)
Subphylum	*Vertebrata*	vertebrates (with backbone and spinal column)
Class	*Mammalia*	mammals
Order	*Cetacea*	whales and dolphins
Suborder	*Mysticeti*	baleen whales (possessing baleen plates instead of teeth)
Family	*Balaenopteridae*	rorquals (possessing pleated throat grooves)
Genus	*Balaenoptera*	finback whales
Species	*Balaenoptera musculus*	Blue whale

In addition to taxonomic classification which identifies an organism, any particular organism also has characteristics which might be worth recording in a database. When you go for an annual checkup, the doctor records your name (your identification) and also your height and weight – traits about you. That blue whale might be up to 100 feet long and weigh 400,000 pounds. If that particular blue whale was observed in the field, an estimate of length was made, and these data were to be entered in a database, both the taxonomy (*Balaenoptera musculus*, blue whale) and also the trait (length = 100 feet) would be necessary to sufficiently describe that biological observation. This distinction between taxonomy and trait sets the context for the ensuing discussion.

3.4. THE DATA CUBE

This distinction in data management approaches arises from (and leads to) differing data dimensionality. Conceptually, a single data value can be thought of as occupying a unique combination of space, time, and variable – where, when, and what was measured. This conceptual approach can be visualized using a data cube (Figure 11) (Maidment 2002).

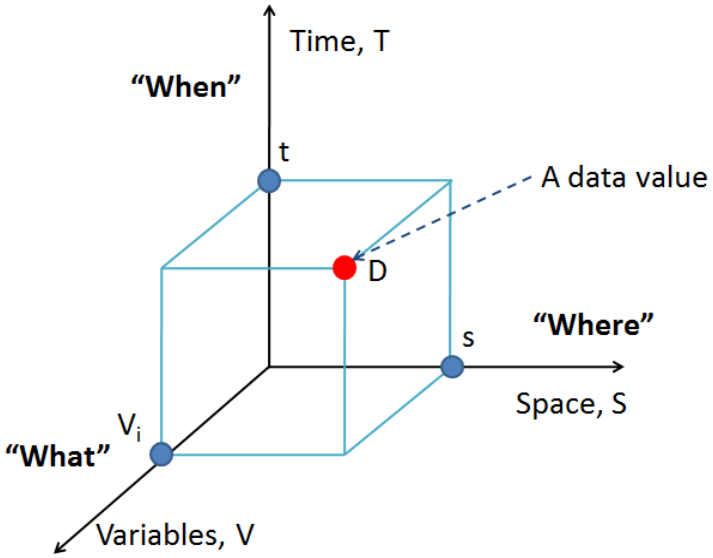


Figure 11. The data cube (Maidment 2002).

A collection or series of multiple data values taken together can be visualized as 'slices' across the data cube. Figure 12a depicts all values across all time at one location; Figure 12b is one variable measured across all time and all space (such as data commonly expressed in raster format); and Figure 12c is all values at one point in time. This is akin to a "time series" – one location, one variable, with a set of observed values through time (a line in the T direction from a particular point in the S,V plane) (Figure 13a); and a "collection" – one location, many variables, with a set of values for a particular time (a line in the V direction originating at a particular point in S,T plane). Similarly, raster layers are commonly used to depict one variable at one point in time across all space (Figure 13b); for example, land surface elevation across a watershed.

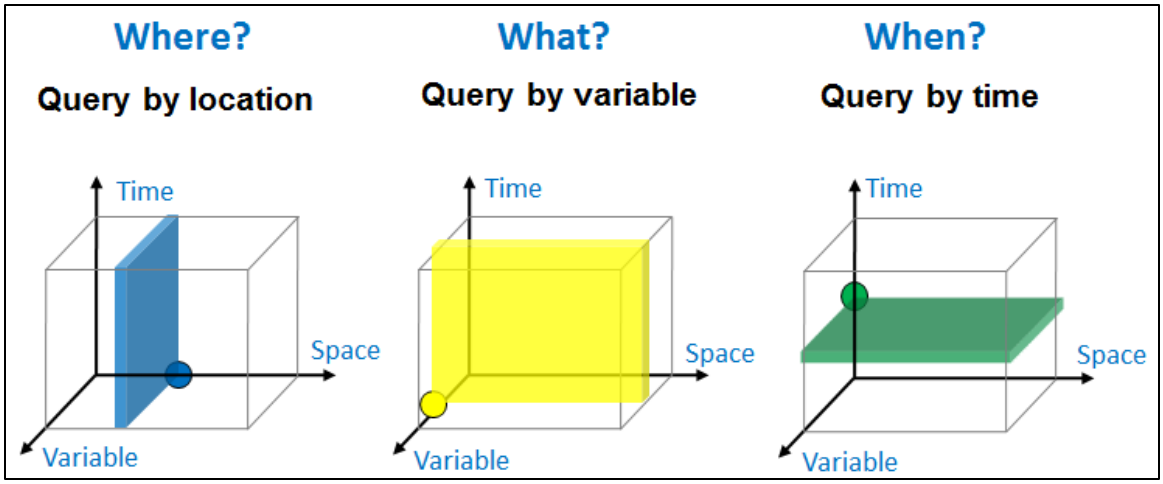


Figure 12. Data cube representations depicting: (a) all values from one station; (b) all values for one variable; and (c) all values at one point in time (sensu Maidment 2002)

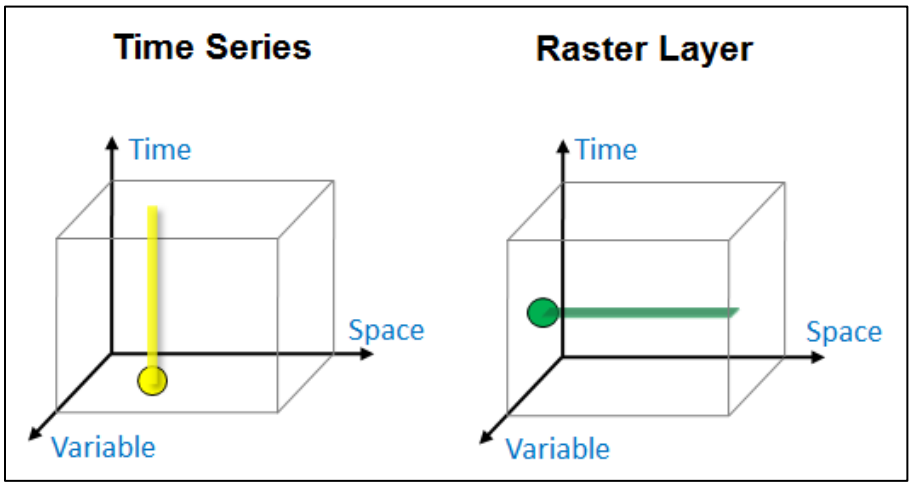


Figure 13. Data cube representations depicting: (a) a time series of values and (b) a raster layer.

Biological data can have many more variables (all the species of the world and numerous traits within each species) but many fewer observations per variable (a handful of samples versus a time series). Biological data tend to be collected by researchers as part of distinct studies, finite in space and time, but sometimes densely-spaced within a collection region. Physical and chemical data for monitoring the water environment are often collected on an ongoing basis, across a wide spatial extent, by public agencies. Occasionally, denser physical and chemical data are collected by researchers on a study basis, often more common with chemical data than physical.

For physical and chemical data, the sample space which defines any particular observation is a function of {x, y, z, t, and v}. x and y represent the location in space, expressed in latitude and longitude, with a known spatial reference system. z is the vertical location, either ignored or treated as an offset from a known datum (such as ground or water surface). t is time, expressed in Coordinated Universal Time (UTC) and the local UTC offset. v is the variable being measured, expressed according to a standardized ontology of variable names. The sample space may be conceptualized using the data cube: the three dimensions of space, time, and variable uniquely define a particular observation (where, when, what) (Figure 14a).

Many of the distinctions between physical, chemical, and biological water data become evident when viewed in the context of the data cube. Physical water data (such as streamflow) tend to have long periods of record, broad spatial extents, and a very limited number of variables measured (Figure 14b). Chemical water data (water quality) tend to have many more variables, a broad spatial extent, and moderately-long periods of

record, although they tend to be much more irregular in time as compared to physical data (Figure 14c). Biological data can have orders of magnitude more variables, but often much shorter temporal extent and much smaller spatial extent. When considering biological data management, the concept of 'variable space' becomes more complicated, however, because variables are needed for both biological species and for descriptors or traits of those species, and these traits may apply to individual species or to collections of species.

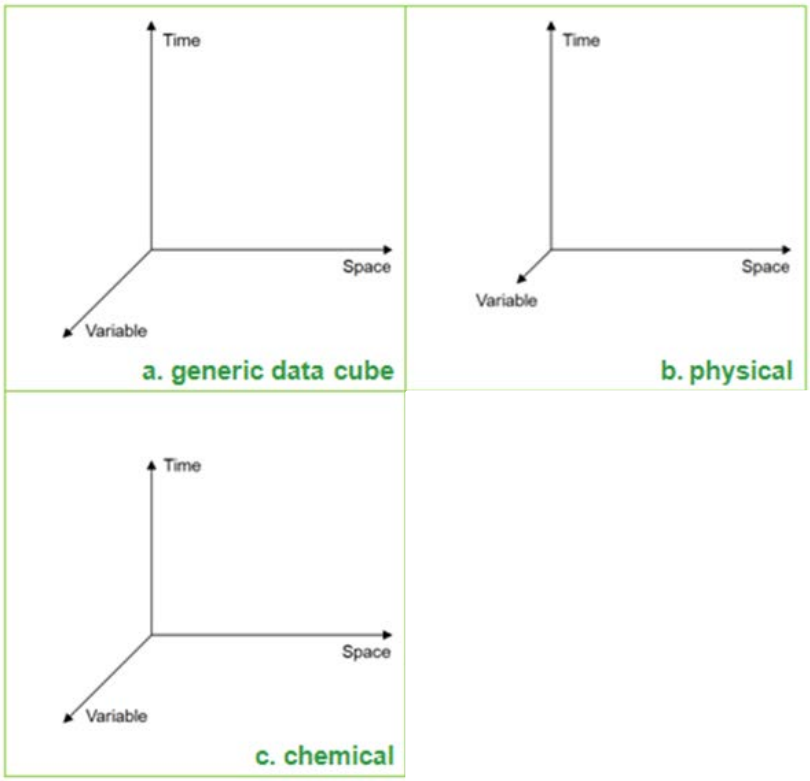


Figure 14. Data cube representations for (a) generic water data; (b) physical water data, which tend to have long periods of record, broad spatial extents, and a very limited number of variables; and (c) chemical water data, which tend to have more variables, a broad spatial extent, and moderate periods of record.

For biological data, the sample space is still a function of {x, y, z, t, and v}, but the variable is now a combination of the taxonomic identification (such as *Genus species*) and the trait (i.e. the measurements, traits, and characteristics of an organism) such as length, biomass, sex, or count (Figure 15).

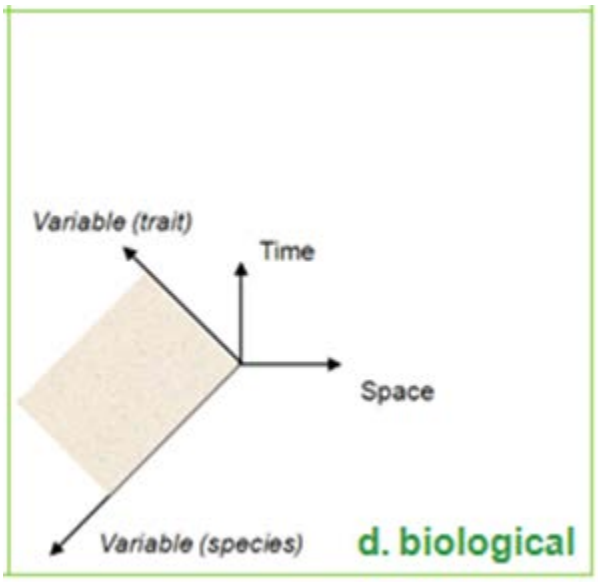


Figure 15. Data cube representations for biological water data, which tend to have much smaller spatial and temporal extents, a relatively small number of traits measured, and a potentially much larger number of taxa observed.

In essence, taxonomy is added to the data cube as a fourth dimension with the existing variable axis taken to mean 'trait.' This is important because taxonomy is often used to index and perform searches on observations within a biological data collection, analogous to date and time in a time series. The 3-D data cube characterizes values

within space and time whereas a biological-specific data model built around the 4-D data cube would also characterize phenomena in 'variable space.' As such, if an ontology of variables is defined to facilitate data discovery it must allow for searching via taxonomy and not only via trait.

Had either taxonomy or trait been solely recorded in the database, it would not be known how long the Guadalupe bass was (if only taxonomy were recorded) or it would not be known what organism was 18 cm in length (if only the trait were recorded). Obviously, both of these solutions are inferior.

An early solution for storing biological observations data considered during this research was to use a variable definition which was a concatenation of taxonomy and trait; for example, *Guadalupe_bass.total_length*. With this variable name, biological observations could be shoehorned into the existing 3-D data cube without the loss of information content exemplified above. This hybrid variable name has two distinct shortcomings, however. First, a new variable must be added to the database for every combination of taxonomy plus trait. If 55 fish species were observed in a sample, and if organism count, average length, minimum length, and maximum length are recorded for each species, the database variable dimension balloons in size to 55 species x 4 traits = 220 variables. This results in a database with a considerable number of null values – computationally inefficient for large databases.

Second, and perhaps more importantly, this hybrid variable system is particularly difficult to query. If a researcher wanted to know everything about Guadalupe bass in a particular dataset, they must first know exactly which traits were measured for the

Guadalupe bass. In the above example, this would be 4 queries. If the researcher wanted to know the count for every species in a sample (such as would be necessary to calculate relative abundance), they must query every species. In the above example, this would be 55 queries.

When taxonomy is added to the data cube as a fourth dimension with the existing variable axis as trait, it becomes possible to easily and efficiently query by taxonomy and/or trait. Databases can be queried by species ("I want to know everything about Guadalupe Bass (*Micropterus treculii)* in the Blanco River.") and also by trait ("What is the relative abundance (the trait) of Bering Flounder (*Hippoglossoides robustus)* observed in the Beaufort Sea?" or "What is the average length (a statistic calculated on the trait "length") of Bering Flounder observed in the Beaufort Sea?") In this sense, there are 'Traits that have Species' and 'Species that have Traits' and both can be easily searched and discovered.

3.5. SEMANTIC MEDIATION

What and how we name the thing being measured is non-trivial and can lead to significant difficulties in the exchange of information. Variations in how data are formatted can be addressed via standardized formats for storage and communication. For example, CUAHSI achieves this syntactic mediation through the use of the Observations Data Model and WaterML web services. Variations in how data are described can be addressed via a standardized translation among languages of variables – semantic mediation. Semantic mediation allows for the communication of data *across multiple systems* – a common terminology allows for information sharing and prevents data duplication.

Some examples: (1) reservoir inflow versus discharge. Both are the volumetric flow of water. The two terms represent the same variable, but are expressed in different terms native to specific domains (reservoir operations versus hydrologic science). (2) ‘Gage height’ versus ‘stage height’ versus ‘stage’ versus ‘water level’ (Maidment 2008). As shown, subtle differences in how an observation is described can lead to significant confusion in data discovery and interpretation. Thus, an ontology is employed – a formal description of the concepts and relationships within a domain. "An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents.” (Gruber 1995)

3.6. ONTOLOGIES

CUAHSI has developed a hydrologic science ontology which uses as a starting point NASA's Global Change Master Directory and incorporates the chemical substances catalogued in the EPA's Substance Registry System; this latter system is also used for chemical characterization by the USGS. The CUAHSI ontology is structured in a stem and leaf pattern, where "Keywords are arranged hierarchically and end in a 'leaf concept' using the same terms as the ODM Controlled Vocabulary…The higher levels of the ontology, which contain many child elements, are only for navigation and are not searchable because the returns would be too large (e.g., "Chemical properties"). Lower levels of the hierarchy are searchable (e.g., "Nutrients") and will return all child concepts." (CUAHSI HIS 2008)

Eleven media are represented in the CUAHSI ontology (e.g.: air, surface water, groundwater, snow, tissue) within three domains layers (physical, chemical, biological) (Figure 16). The biological domain of the CUAHSI ontology was developed and refined as part of this proposed work and currently includes taxonomic classification at the Family, Genus, and Species levels within six subgroups (benthic, fish, macroinvertebrates, nekton, phytoplankton, and zooplankton) (Maidment 2009).

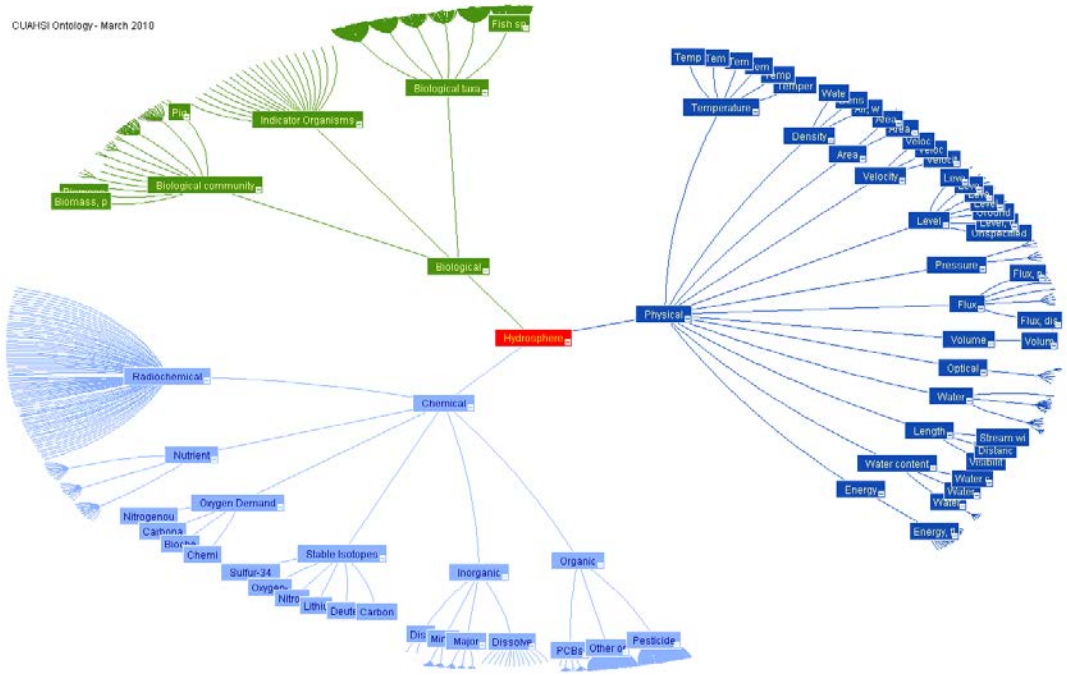


Figure 16. The CUAHSI hydrologic data ontology.

The CUAHSI Controlled Vocabulary master list is comprised of 13 sublists – CensorCode, DataType, GeneralCategory, SampleMedium, SampleType, SiteType, SpatialReferences, Speciation, TopicCategory, Units, ValueType, VariableName, and VerticalDatum. The CV refers to specific dimensions within an ODM database, *for all data and metadata*, such as sample medium, variable name, spatial reference, and units, and its purpose is to standardize how data are described (i.e., to provide semantic and syntactic mediation). Relatedly, the CUAHSI ontology standardizes how *variables* are uniquely described and provides a single term for data queries. The CUAHSI VariableNameCV and the ontology have very significant overlap, and data publishers

ideally will tag their variables onto the ontology using the Controlled Vocabulary terminology when they register new data services. This tagging is performed manually now but could be performed automatically in the future.

Another data discovery improvement currently under development is faceted search, where the number of metadata dimensions by which a search can be performed is expanded. In faceted search, a user selects a facet upon which to search (spatial extent, time window, variable category or variable), and the search interface dynamically updates to show the trimmed-down results field. Faceted search is common today in online shopping and travel sites, where price, customer rating, size, and style are example facets. In the CUAHSI HIS, faceted search will likely include sample medium, sampling organization, space, time, and variable, all accessed via a map interface (Bedig 2011).

The biological component of the CUAHSI ontology has at its origin a study made by the author of the biological data stored within the TCEQ Regulatory Activities Compliance Systems (TRACS) Surface Water Quality Monitoring (SWQM) database. "STORET parameter codes within TRACS were divided into groupings of biologic and ecologic significance (Table 4). SQL Queries were then performed in Microsoft Access to extract appropriate data and statistics, and the results of these queries were summarized (Table 5). From these analyses, it was confirmed that TRACS contains relatively little data (number of records) specific to biology, but a large proportion (over 55%) of the codes in TRACS are dedicated to biologic data." (Hersh 2007)

Table 4. Groupings of EPA STORET parameters developed for analysis of the TCEQ Surface Water Quality Monitoring database (Hersh 2007).

Category	Description
Site and Sample	Including: sampling effort, methods, and equipment; substrate, channel geometry, geomorphology, streamflow, cover, vegetation, watershed, aesthetics, and weather
Benthic Macroinvertebrates	Animals without backbones which live all or part of their lifecycle in or near the bottom of freshwater systems. Including: Platyhelminthes (flatworms), Annelids (worms, leeches), Arthropods (mites, insects, crustaceans), and Mollusks (clams, mussels, snails)
Fish	Vertebrate cold-blooded animals that live their entire lives in water, breathe by means of gills, and move by means of fins (with some exceptions)
Phytoplankton	Microscopic, free-floating or suspended plants and algae which have movement depending on currents and are primary producers
Zooplankton	Microscopic animals capable of movement and are secondary producers. Including: crustaceans and rotifers, diatoms, dinoflagellates, and copepods.
Nekton (non-fish)	Free swimming organisms, exclusive of fish as defined above. Including: Decapods (shrimp, prawns, crayfish, crabs), jellyfish, squid, turtles, frogs, alligators
Macrophytes	Large vascular aquatic plants, growing in or near water that are either emergent, submergent, or floating. Including: cattails, rushes, arrowhead, waterlily

Hierarchically, the CUAHSI ontology has “Hydrosphere” at its core, then the three domains (physical, chemical, biological), then multiple stems, many branches, and numerous leaves, also known as “leaf concepts.” The procedure for adding new leaves to the ontology (i.e., new variables) is defined by the CUAHSI HIS team and includes requesting new additions to the controlled vocabulary (Maidment 2009).

Table 5. Summary of biological data in TCEQ Texas Regulatory and Compliance System Surface Water Quality Monitoring database. (Hersh 2007)

Category	Data Values		Variables	
	Count	% of TRACS	Count	% of TRACS
TRACS SWQM	7,591,675	100.00%	4,412	100.0%
Site and Sample	184,935	2.44%	82	1.9%
Benthic Macroinvertebrates	49,402	0.65%	1,323	30.0%
Fish	32,710	0.43%	311	7.0%
Phytoplankton	10,099	0.13%	371	8.4%
Zooplankton	10,344	0.14%	266	6.0%
Nekton (non-fish)	2,942	0.04%	31	0.7%
Macrophytes	449	0.01%	54	1.2%
Total, Biologic Data in TRACS	**290,881**	**3.83%**	**2,438**	**55.3%**

Under the Biological domain of the CUAHSI ontology are three stems: Biological taxa, Indicator organisms, and Biological community. The Biological taxa stem is where the six categories of TCEQ SWQM data are represented (excluding Site and Sample) (Figure 17), and each of these categories in turn includes tens to hundreds of taxa as species. The ontology includes leaves only for those species currently represented in the data cataloged at HIS Central but the stem-and-leaf structure of the ontology allows for the additional leaves as needed to accommodate new taxa. The biological domain includes three stems: biological taxa, indicator organisms, and biological community. Here is an example to illustrate the hierarchical nature of the CUAHSI ontology. If a

researcher had collected data on the how many fish in a seine haul sample have a diet consisting primarily of other fish, she would tag those data as shown in Table 6.

Table 6. An example to illustrate the hierarchical nature of the CUAHSI ontology.

Ontology Level	**Value**
Core	Hydrosphere
Domain	Biological
Stem	Biological Community
Branch	Fish
Leaf Concept	% of individuals as piscivores, fish

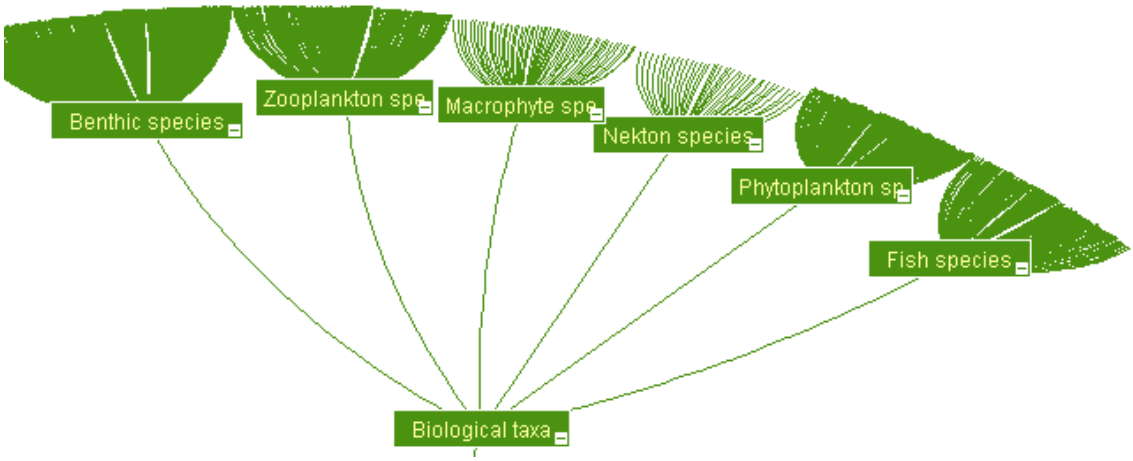


Figure 17. Biological taxa hierarchy of the CUAHSI ontology. (CUAHSI HIS 2011)

The Biological community stem includes a number of metrics commonly used in the characterization of a community of organisms and in the assessment of ecological

health (Figure 18). It includes metrics common to the development of the Index of Biological Integrity (IBI) for fish and macroinvertebrates (Karr 1981) (Table 7) plus metrics for planktonic biomass, for assessing fish kill severity, and for chlorophyll- and non-chlorophyll-based pigments (CUAHSI HIS 2011).

Table 7. Metrics used in determining the Index of Biotic Integrity (Linam et al. 2002).

Metric	Karr et al.	24	25,26	27,29,32	30	31	33,35	34
Total number of fish species	X	X	X	X	X	X	X	X
Number of darter species	X							
Number of native cyprinid species		X	X	X	X	X	X	X
Number of benthic invertivore species		X		X	X		X	X
Number of benthic species						X		
Number of sunfish species	X	X	X	X	X	X	X	X
Number of sucker species	X							
Number of intolerant species	X	X			X		X	X
% of individuals as green sunfish	X							
% of individuals as tolerant species (excluding western mosquitofish)		X		X	X	X	X	X
% of individuals as omnivores	X	X	X	X	X	X	X	X
% of individuals as insectivorous	X							
% of individuals as invertivores		X	X	X	X	X	X	X
% of individuals as piscivores	X			X	X	X	X	
Number of individuals in sample	X							
Number of individuals per unit effort		X	X	X	X	X	X	X
% of individuals as hybrids	X							
% of individuals as non-native species		X	X	X	X	X	X	X
% of individuals with disease or other anomaly	X	X	X	X	X	X	X	X

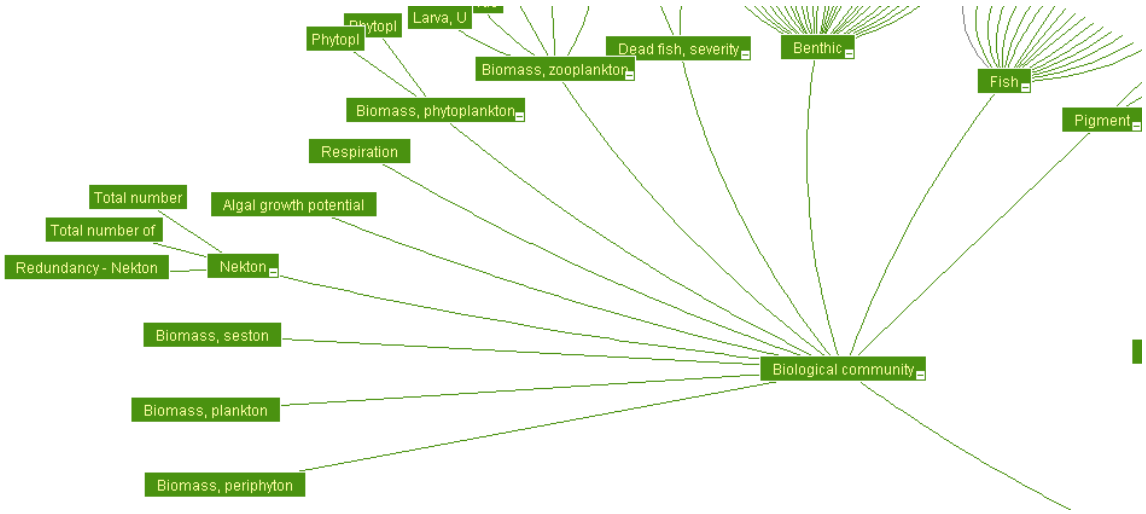


Figure 18. Biological community hierarchy of the CUAHSI ontology. (CUAHSI HIS 2011)

3.7. BioODM

CUAHSI has developed and refined the Observations Data Model (ODM) for the storage and retrieval of series-based hydrologic observations in a relational database. As discussed, some incompatibilities arise when trying to employ the ODM for biological information due to the collections-based nature of these data. Thus, a relational data model was developed for biological observations in the aquatic environment – BioODM.

BioODM is designed to directly associate the water environment (i.e. habitat) with its inhabitants. Conceptually, BioODM differs from the CUAHSI ODM via: (1) its explicit incorporation of taxonomy and habitat, (2) its treatment of sampling methodology, (3) its reliance on multi-dimensional variable space, (4) its linkage to

documents and other knowledge products, (5) its partnership with geographic data stored within a GIS, and (6) its structure which places the focus on the organism(s) observed (Figure 19 and Figure 20). For comparison, version 1.1 of the CUAHSI ODM has no capacity to simultaneously incorporate taxonomy and traits, has limited sampling methodology resolution (e.g. no ability to address sampling effort or gear size), and has a focus on a single observation indexed within a time series. As such, the BioODM presented here is a conceptual, idealized version of a relational data model for aquatic biology data.

All that being said, the existing CUAHSI ODM offers considerable flexibility and adaptability for the storage of physical and chemical data for the water environment, however, and it has the distinct advantage of having tools developed which facilitate its use: (1) the ODM Data Loader software to input data into the XML schema, (2) the Time Series Analyst to graphically view the data and perform limited statistical analysis, and (3) the WaterML web language and the WaterOneFlow web services to communicate the data. As a result of the support infrastructure already in-place for the CUAHSI ODM, there is a significant strategic advantage in adapting and refining the structure of the existing CUAHSI ODM rather than starting from scratch.

In light of this recognition, a pragmatic, middle-ground BioODM was developed here which is designed to leverage as much as possible of the CUAHSI ODM infrastructure and institutional support while adding the critical components necessary for the adequate storage and representation of biological data for the water environment. This compromise BioODM added a Taxonomy table to the CUAHSI ODM and a TaxaID

Foreign Key to the central DataValues table. The Taxonomy table includes a unique TaxaID as its Primary Key; optional fields for the Family, Genus and/or Species of an organism; and an optional field for Comments related to the taxonomic identification and/or classification (Table 8).

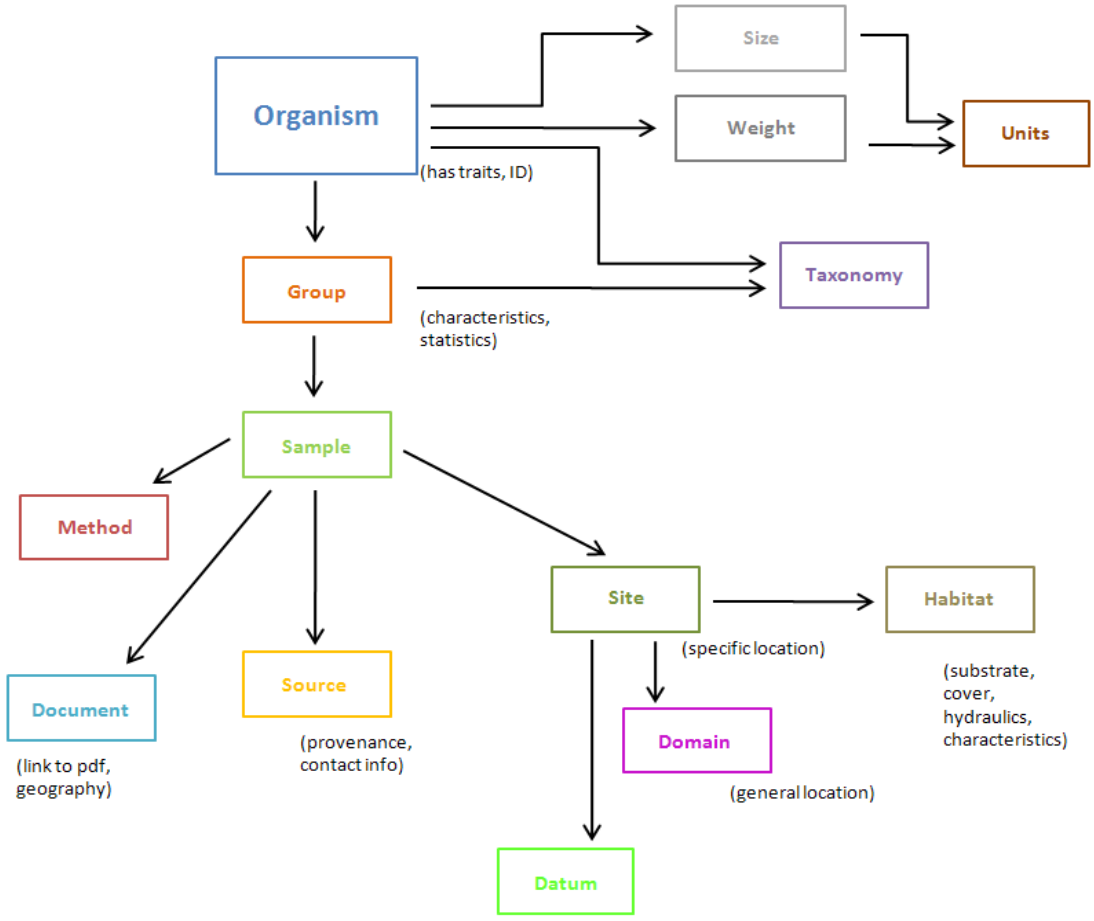


Figure 19. Schematic representation of the BioODM, version 1.2.

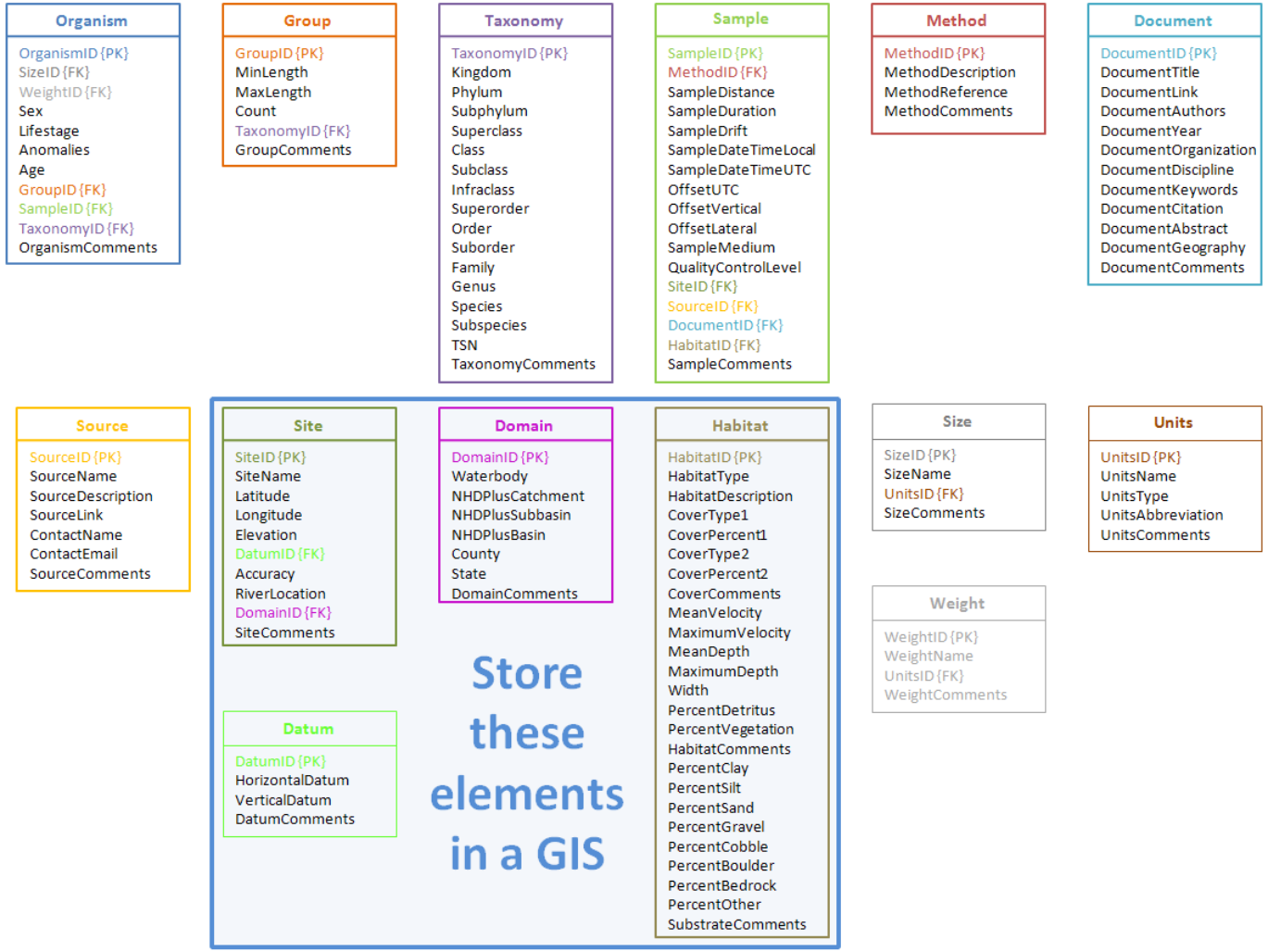


Figure 20. BioODM table specification, version 1.2.

Table 8. ODM Taxonomy table fields and specifications.

Field Name	Data Type	Description	Example	Constraint
TaxaID	Integer, Identity	Unique integer identifier for each taxonomic classification	42	Mandatory; Unique Primary Key
Family	Text (50 char)	Scientific family name	*Centrarchidae*	Optional
Genus	Text (50)	Scientific genus name	*Micropterus*	Optional
Species	Text (50)	Scientific species name	*salmoides*	Optional
TaxaComment	Text (256)	Comments related to the taxonomic identification and/or classification	-	Optional

As discussed, the taxonomic classification represented within the Taxonomy table is performed according to the CUAHSI variable ontology which has been derived from the Integrated Taxonomic Information System (ITIS) species classification, the authoritative taxonomic catalog for the United States (http://www.itis.gov/). In that regard, ITIS can be thought of as the controlled vocabulary for species identification, and semantic mediation for taxonomic identification is accomplished internal to the ITIS program. The benefit of used a standardized, hierarchical taxonomic classification include the ability to aggregate data searches and analyses up the taxonomic chain. For example, organisms can be stored in the database at the species level but can be queried at the family or genus level. ITIS resulted from an interagency partnership formed to address deficiencies identified by the White House Subcommittee on Biodiversity and Ecosystem Dynamics in federal systematics efforts, especially in the organization, access,

and standardized nomenclature of data necessary to support ecosystem management and biodiversity conservation (ITIS 2012). The ITIS partnership founding members include:

- Department of Commerce – National Oceanographic and Atmospheric Administration (NOAA);
- Department of Interior (DOI) – US Geological Survey (USGS);
- Environmental Protection Agency;
- US Department of Agriculture (USDA) – Agricultural Research Service (ARS) and Natural Resources Conservation Service (NRCS);
- Smithsonian Institution – National Museum of Natural History;

and additional current partners include:

- Department of Interior – National Park Service (NPS) and US Fish and Wildlife Service (USFWS);
- NatureServe;
- Comisión Nacional para el Conocimiento y Uso de la Biodiversidad of Mexico (CONABIO); and
- Agriculture and Agri-Food Canada (ITIS 2012).

3.8. DATA THEMES

ODM databases are typically organized by data source/provider; that is, for each agency and for each observations network that supplies water data, there is a separate ODM database. Unfortunately, this isn't the most desirable format to many users – it is

common to seek data across all data providers for one variable, one related group of variables, or for a specific geographic extent. These sets of geospatial or observations data grouped together for some purpose are called themes, and they are a promising avenue for biological water data management. Themes are common in a GIS sense, and have recently been extended to an observations sense. (CUAHSI 2010) Themes may be implemented via a thematic dataset catalog – a feature that contains geospatial information, summary data, and the information required to call a web service to retrieve time series data for each site (Seppi 2009; Whiteaker 2009).

A theme for Texas salinity was developed which included data for one variable (salinity) across multiple data providers (TCEQ, TWDB, and TPWD), and across a wide spatial extent (Texas rivers and coast) (Figure 21).

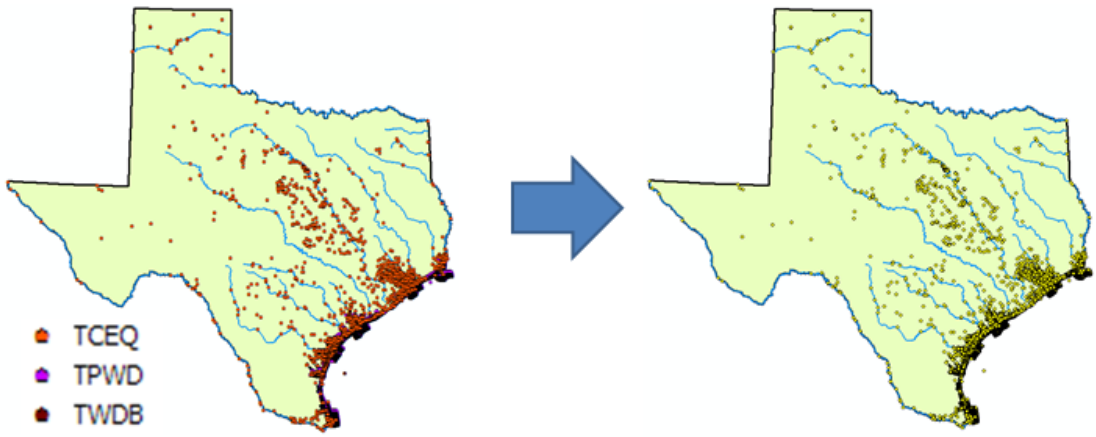


Figure 21. Example of a 'Texas salinity' data theme, where observations data from multiple data providers (TCEQ, TPWD, and TWDB) are merged into a unified data theme for salinity across the State of Texas.

The themes concept could be extended to support data discovery for the disciplines of the Texas environmental flows program: hydrology, water quality, geomorphology, and aquatic biology (Figure 22).

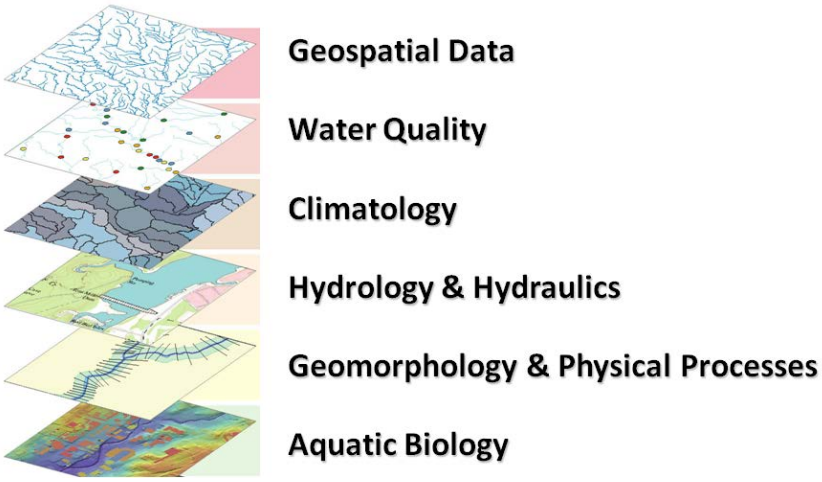


Figure 22. Thematic representation of the Texas environmental flow program disciplines.

3.9. CONCLUSIONS

As has been shown in this chapter, biological observations of the water environment and thus their associated data differ strongly from physical and chemical observations. A 4-D data cube was developed to accommodate a primary difference whereby the variable space of the traditional data cube is re-envisioned to include both the organism's taxonomy and also its traits. An ontology was developed for use in the

CUAHSI Hydrologic Information System based on an analysis of approximately 30 years of data from the TCEQ TRACS database using the EPA STORET parameters. A BioODM data model was developed in both a conceptual and practical variation. Collectively, these modifications serve to adapt the existing CUAHSI Observations Data Model for use with biological observations of the water environment.

Because of the 4-D data cube innovation, databases can be queried by species ("I want to know everything about Guadalupe Bass (*Micropterus treculii)* in the Blanco River.") and also by trait ("What is the relative abundance of Bering Flounder (*Hippoglossoides robustus)* in the Beaufort Sea?") In this case, the Integrated Taxonomic Information System (ITIS), the authoritative taxonomic catalog for the United States, was used to provide a standardized species nomenclature and thus to help rein-in the vastness of biological data.

Chapter 4: Managing Arctic Marine Observations Data

4.1 INTRODUCTION

The Arctic is changing. Temperatures are warming and the minimum sea ice extent is retreating (Pachauri and Reisinger, 2007). Changes in the presence and condition of sea ice are stressing some ice-dependent species such as polar bears (CFR 2010). On shore, the yield of the Prudhoe Bay oil field has diminished and the Trans-Alaska Pipeline is operating below capacity (API 2009). America's continued thirst for oil and gas has led to an increased desire to explore new offshore sources, including the outer continental shelf regions of the Chukchi and Beaufort Seas off the northwest and north coasts of Alaska. In 2008, the Minerals Management Service (now the Bureau of Ocean Energy Management) generated $2.6 billion in high bids for 488 blocks under Lease Sale 193 (MMS 2008, MMS 2008b). The Chukchi Sea Offshore Monitoring in Drilling Area: Chemical and Benthos (COMIDA CAB) project was initiated in 2008 to be a robust, comprehensive effort to characterize the lease area biota and chemistry, to conduct a baseline assessment of the continental shelf ecosystem via ship-based physical, chemical, and biological sampling of the benthos, and to develop a workable food web model.

The COMIDA CAB effort involves seven Principal Investigators hailing from five universities and one Contracting Office Representative. Over two field seasons aboard the R/V Alpha Helix (summer 2009) and the R/V Moana Wave (summer 2010) in

the northeastern Chukchi Sea, the project team collected diverse observational data from multiple instruments and sensors, in varying sample media, across varying spatial and temporal scales, in the broad disciplines of physical, chemical, and biological oceanography. In all, a total of 48 stations were occupied in 2009 and 44 in 2010 including 27 stations which were reoccupied for quality control and time series comparative purposes (Figure 23). One operational goal for this project is to establish an environmental baseline so that "undisturbed" conditions can be described prior to the commencement of oil drilling activities. This necessitates the compilation of information from the project into a database synthesized in a uniform way across the study area rather than having just the original investigator files.

As can be expected from such a multi-disciplinary effort, data management is an important and challenging task. The COMIDA CAB project includes a dedicated, ship-board data manager to provide real-time, field-based data services and Geographic Information System (GIS) support. Project data management is accomplished via the SQL/Server relational database and the Observations Data Model (ODM) relational database schema. The ODM originates from the Consortium of Universities for the Advancement of Hydrologic Science – Hydrologic Information System (CUAHSI HIS), a National Science Foundation-supported cyberinfrastructure project for the hydrologic sciences, used extensively for storing observations of the physical, chemical, and biological components of the water environment (Maidment 2009, Horsburgh et al. 2008).

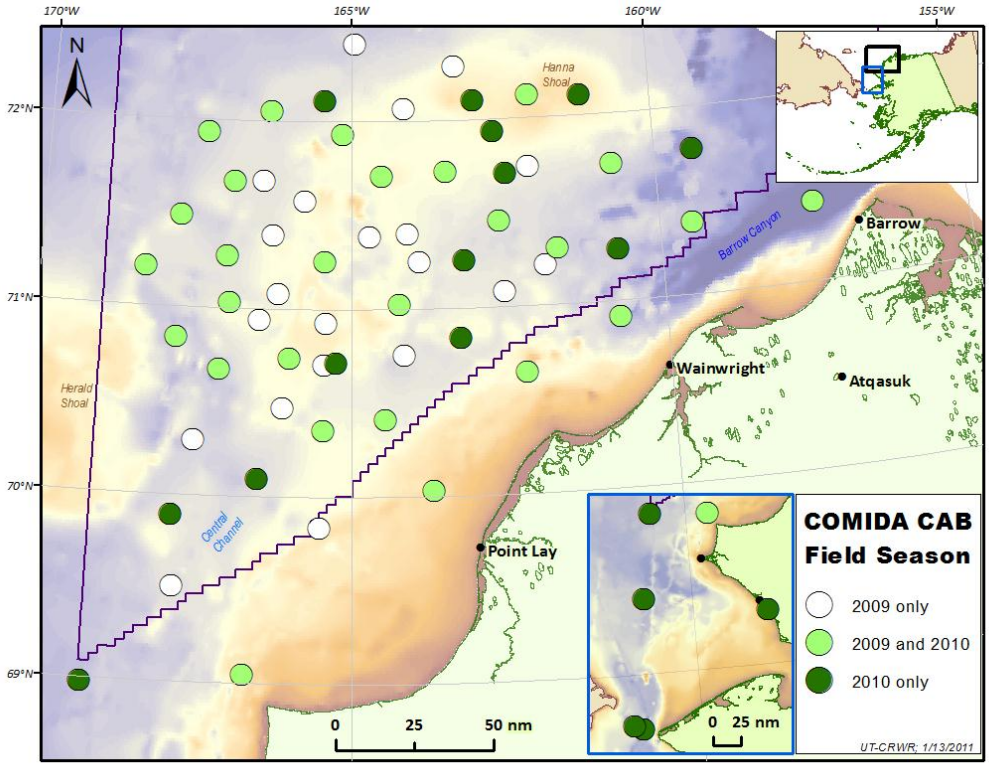


Figure 23. Stations occupied during the 2009 and 2010 COMIDA CAB field seasons in the northeastern Chukchi Sea, Alaska.

But actively managing data during the project isn't enough. The size, scope, and interdisciplinary nature of this project results in a wealth of information and represents a significant research investment. Effective project data management must include public outreach, data sharing, and data archiving both during and after the life of the project. As such, a secure, web-based system was developed for observational data storage (via the Integrated Rule-Oriented Data System (iRODS), geographic data storage (via the ArcGIS Online community), document sharing, and public outreach (Rajasekar et al 2009, Rajasekar et al 2006).

Thus, the objectives of this chapter are, broadly: to present an approach to making observations of the ocean environment, to put forth a methodology for organizing and storing these observations, and to offer various avenues for communicating scientific results widely via the use of open standards.

This chapter addresses the challenges of biological data management in a relatively narrow academic study. It presents the adaptation of the CUAHSI Observations Data Model for application with physical, chemical, and biological oceanographic data – a new extension of the CUAHSI Hydrologic Information System, thus bringing hydroinformatics into the oceanographic realm.

4.2 The Nature of Oceanographic Data

Ocean science is multi-disciplinary and conducting ocean research is logistically complex. While remote and satellite-based sensing are common in physical oceanography, chemical oceanography and marine biology largely require in-situ field sampling. Researchers across the globe are connected by a common interest in many of the same questions in many of the same oceanic regions so it is important that data from individual cruises are stored and made permanently accessible. Many parallels exist in collecting, organizing, and storing data for freshwater and marine ecosystems, and though the organisms observed may differ greatly from land to sea, many of the same collections-based principles apply. Some differences exist, however, with respect to biological data management for freshwater versus marine systems.

From a geographic perspective, freshwater systems focus on a waterway (e.g. river, stream, creek, lake, pond, or reservoir) with observations made at point locations, such as at gaging stations, grab sample locations, etc. Marine systems often feature a much broader physical area with sampling performed in a much more spatially dispersed fashion, but due to the perception of higher spatial homogeneity within the marine environment, marine observations are often interpolated across a much larger physical area than freshwater observations. Marine observations are either made at a point (on the water surface, in the water column, or on the seafloor) or along a moving track (such as a ship's path).

A sampling 'point' in an oceanographic study may not actually be that – wind and waves may push a vessel off-station resulting in a positional accuracy much less than that for a 'fixed' point (such as a gaging station) in a land-based freshwater study. Although the use of towable acoustic instrumentation is increasing in riverine sampling, marine sampling has traditionally featured more data collection from moving tracks: a ship path, tagged organisms, drifters, and remotely-operated vehicles (ROVs) are a few examples. Lastly, oceanography often features a wider variety of media sampled across a wider array of vertical zones – atmosphere, water surface, water column, epibenthos, benthos, and subfloor.

The character of the data from the COMIDA project differs fundamentally from the majority of typical hydrologic information such as for precipitation or streamflow with regard to the time-domain, or, more specifically, the lack thereof. While traditional water data visualized using a data cube include a time dimension, the COMIDA project

data really only includes a time stamp – that is, when one-time samples were collected, rather than being regularly recorded through time. This distinction is not unique to COMIDA as it is shared by other environmental sampling studies, but is important nonetheless in the consideration of project data management.

Just as data from a terrestrial environmental flows case study can be conceptualized by discipline (hydrology, water quality, geomorphology, and aquatic biology) and by data provider (state agency, EPA, USGS, etc), so too can COMIDA data be conceptualized by discipline (physical, chemical, or biological oceanography) and by data provider (in this case, project Principal Investigators instead of agencies). So even though data from the terrestrial case study comes from *surveys* and data from the Alaska case study comes from *studies*, the conceptual approach to managing water data for each study has much in common (Figure 24).

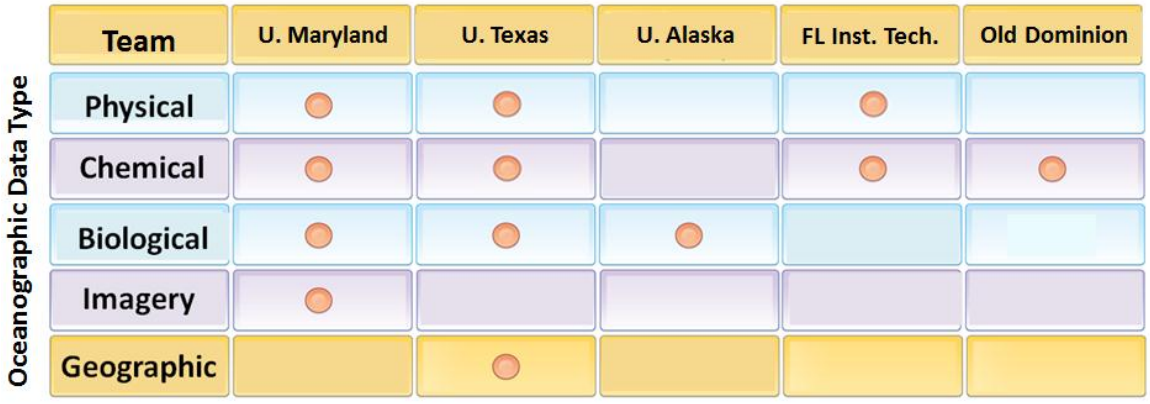


Figure 24. Thematic organization of COMIDA CAB data by Principal Investigator institution and by data type.

4.3 OBSERVING THE OCEAN ENVIRONMENT

4.3.1 Basemap Development

The study area extends from approximately 65º N to 72º N and from 169º W to 157º W. A basemap was developed for the study area based on bathymetric data from the NOAA National Geophysical Data Center. The ETOPO1 1-Arc Minute Global Relief Model (Figure 25) integrates land topography and ocean bathymetry from numerous global and regional data sets (Amante and Eakins 2008).

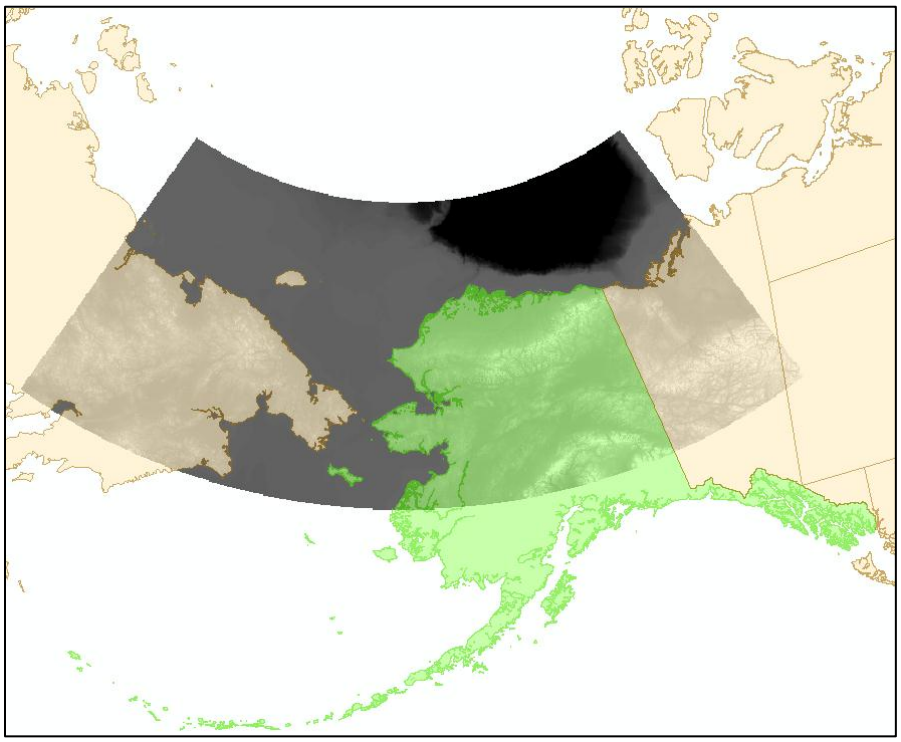


Figure 25. ETOPO1 1-Arc Minute Global Relief Model (Amante and Eakine 2008).

Coastlines, cities, and political boundaries are added for spatial orientation and oil and gas wells, Bureau of Ocean Energy Management Lease Sale Area 193 information, existing moorings, and previous sampling locations are added to provide a context of former and current energy development activities (Figure 26).

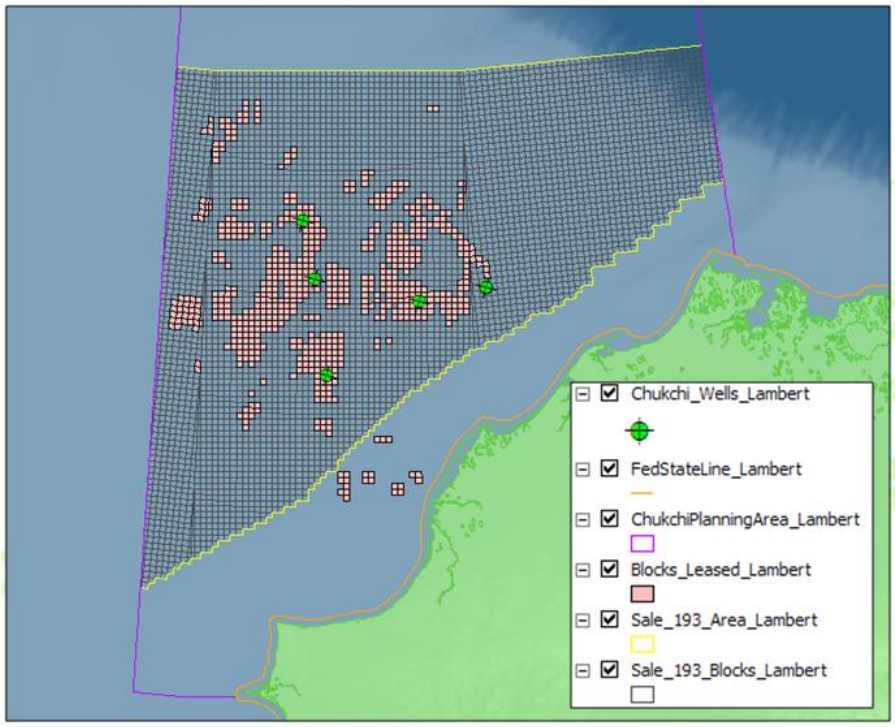


Figure 26. Chukchi Sea basemap data for the Bureau of Ocean Energy Management's Lease Sale Area 193.

4.3.2 Sampling Design

Station locations were determined via two methods for random yet even distribution: (1) a general randomized tessellation stratified design (GRTS) in the core project area (Figure 27), and (2) a spatially-oriented, nearshore-to-offshore, south to north grid overlaying the GRTS design. This arrangement allowed for putting the core station sites in a spatial grid. Of the 30 GRTS stations, 10 were chosen as overlap stations to cross-calibrate and provide QA/QC based on replicate benthic samples. The GRTS design was based on the approach employed by the US Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) (White et al 1992). The grid stations were positioned to provide insight into upstream and downstream conditions with some stations outside the lease area as control sites. Twenty seven stations were sampled in both field seasons to provide a time-comparison of benthic and water column parameters and for quality control purposes.

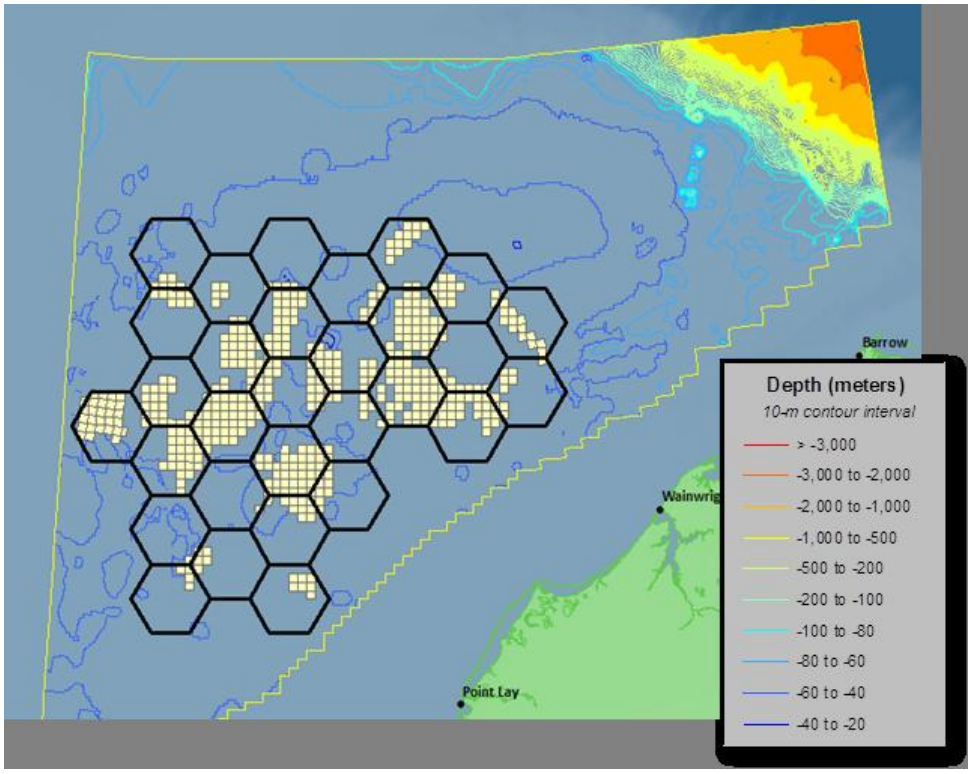


Figure 27. General randomized tessellation stratified (GRTS) design for COMIDA station selection.

4.3.3 Data Collection

Observations were made of the water column, sediments, epibenthos and benthos. During the 2009 field effort, 270 sampling events occurred totaling 142 hours of sampling time with events such as: epibenthic trawls, data sondes, light meters, discrete-depth water column pumping, double van Veen grabs, single van Veen grabs, HAPS sediment cores, box cores, phytoplankton nets, zooplankton nets, and benthic camera deployment. During the 2010 field effort, 273 sampling events occurred over 117 hours of sampling time with similar equipment deployed. Project data collected includes

physical, chemical, and biological observations and the associated geographic data plus video and still imagery. A summary of the diverse data collected is shown in Table 9.

Table 9. Examples of the types of data collected in various sample media.

Water Column	Epibenthos	Sediment
Surface & subsurface PAR	Community composition	Hydrocarbons
Chlorophyll a	Abundance, biomass, population size structure	19 anthropogenic metals
POC & POM	Organic contaminants	Cesium and lead dating
Zooplankton	Nutrients, stable isotopes	TOC, POC, nutrients
Phytoplankton	Caloric content	Sediment chlorophyll
Hydrographic profiles	Oxygen consumption	Benthic infauna
Turbidity, TSS, nutrients	Nutrient flux experiments	Biomarkers
Trace metals	Qualitative video habitat survey	Grain size distribution
Fish toxicology		Oxygen uptake experiments
Birds & marine mammals		

4.4 Organizing and Storing Ocean Observations Data

4.4.1 Data Management Workflow

It is the goal of the COMIDA CAB project team and a requirement of the Bureau of Ocean Energy Management (BOEM) contracting procedure that all project data be preserved in the public domain. However, adequate data and metadata management can be a cumbersome and time-consuming task for project scientists unfamiliar with best practices for good data stewardship. In recognition of these dual interests, a data management workflow was developed for the COMIDA CAB project seeking to minimize the burden to project Principal Investigators yet providing sufficiently

described data such that it may be discovered, accessed, and used by others in the future. The workflow is described below and is depicted in Figure 28.

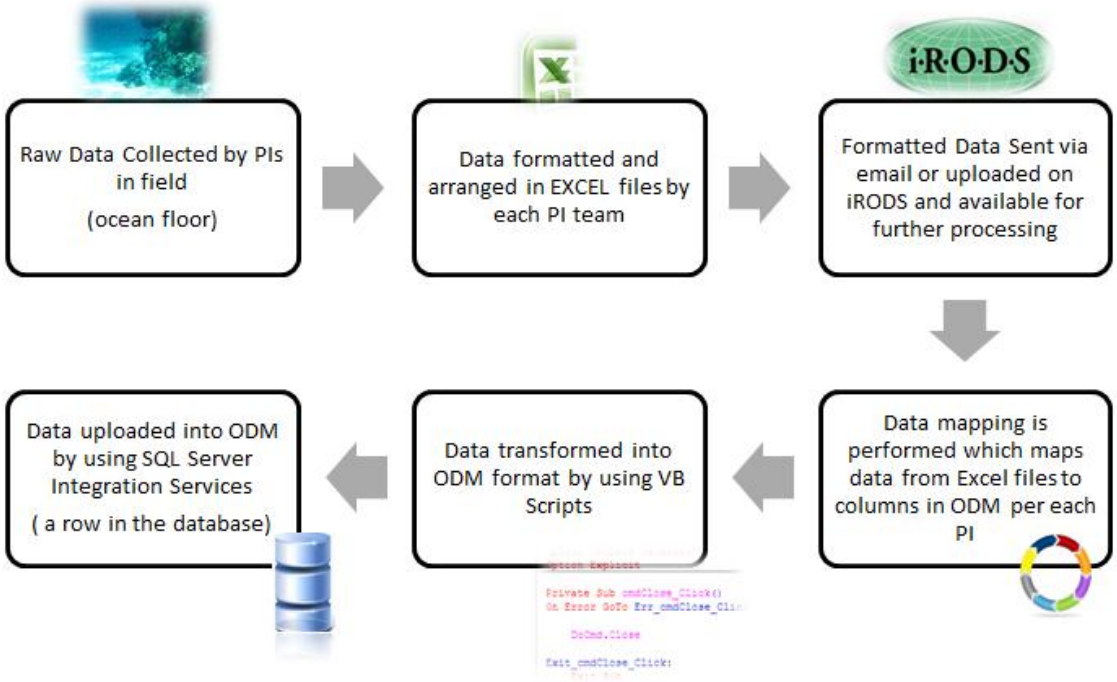


Figure 28. The COMIDA CAB project data management workflow.

Data are collected by PIs while aboard the research vessel or from samples brought back to the laboratory. Data are recorded by PIs in their native, traditional format in Microsoft Excel, a widely used platform which requires no specialized database knowledge. Data spreadsheets are uploaded to a secure online document-sharing system, in this case the Integrated Rule-Oriented Data System, which has been created for the project team and is password-protected. This security is afforded since, according to contract requirements, access to preliminary project data was limited to project

participants and was provided via a log-in interface. The data have now been quality-controlled and approved for release, so project data and analytical products have now become publicly-available.

At this point, the data team 'takes over' management of the data. However, a chain-of-custody is established at the data handoff and this custody signature persists throughout the data workflow, from loading the database all the way through to archiving; the chain-of-custody is described further below. Customized scripts are developed using Visual Basic which establish a 'template' for each PIs data format. These templates serve to map the PIs data to the ODM data model, from the PIs personal terminology to a standardized ontology of variables and controlled vocabularies for metadata.

Since templates are developed for each PI and each data type, repeat and/or revised submissions may be loaded simply and efficiently into the project database; this feature is convenient since the project included two years of field sampling with many similar observations made year-to-year. Finally, data are loaded into the ODM using the CUAHSI ODM Data Loader (http://his.cuahsi.org/odmdataloader.html). This data management workflow may be characterized as an Extract-Transform-Load (ETL) procedure: data are extracted from diverse file formats, transformed into a standardized data structure, and loaded into the project database.

In all, the 2009 and 2010 COMIDA CAB field efforts yielded a database of 510,405 data values. Of these, 474,129 (93%) were derived from sonde profiles and 36,276 (7%) were from non-sonde samples of the sediment, epibenthos, and water

column. The data sondes used in this project are akin to the conductivity, temperature, and depth (CTD) samplers commonly used in oceanographic research. These data values represent 301 variables measured at 65 sites and originating from 26 different source files. The biological observations represented 519 distinct taxa. In this sense, the 4-D data cube has axes with length 301 (variables), length 65 (space), length 519 (taxa), and with the time axis reflecting 1 to 2 measurements made at each station depending on whether the station was resampled in the 2010 field season (Figure 29).

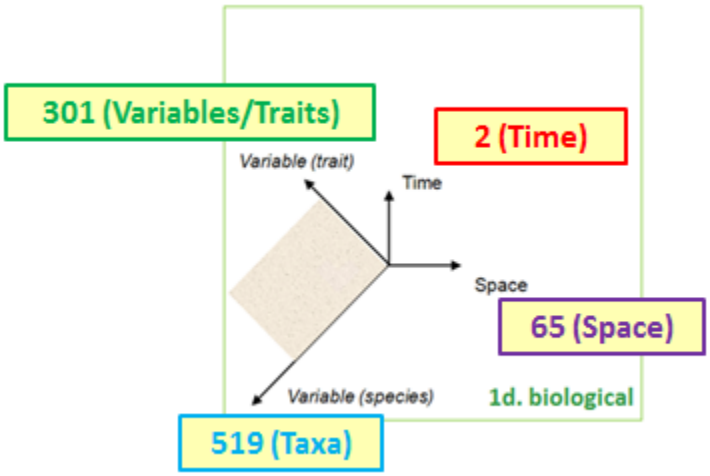


Figure 29. Axis lengths for the 4-D data cube representing the dimensions of the COMIDA CAB project database.

Of the 301 variables represented in the COMIDA CAB project database, 11 (3.7%) are physical in nature, 244 (81.1%) are chemical, and 46 (15.3%) are biological.

4.4.2 Data Model Adaptations

When originally developed, the CUAHSI Observations Data Model was viewed as an extensible "core data model," tolerant of and suitable for customization and adaptation (Horsburgh et al 2008). The ODM has traditionally been used to manage physical and chemical observations of the water environment, so the biological observations data here necessitated modifications to the data model, data loader, and the associated Controlled Vocabulary. Three of the seven COMIDA CAB Principal Investigators collected some type of biological oceanographic data – benthic and epibenthic abundance, biomass, density, diversity, and taxonomy. The ODM is structured as a "star schema" meaning that each data value has primacy and the associated metadata are linked to the primary data value using database relationships. In the case of the biological data represented here, the trait or descriptive characteristic is the data value (abundance, biomass, etc) and thus the taxonomic identification must be accomplished elsewhere in the data model. To accomplish this, a TaxaID foreign key was established within the ODM Data Values table which links to a newly-created Taxonomy table. The Taxonomy table is built upon a ten-level hierarchical taxonomic classification system and includes non-mandatory attribute fields for hierarchical classification along with mandatory inclusion of the Taxonomic Serial Number (TSN) obtained from the Integrated Taxonomic Information System (ITIS, http://www.itis.gov/), the authoritative taxonomic catalog for the United States (Table 10).

Table 10. Attributes of the new ODM Taxonomy table.

Field Name	Data Type	Description	Example
TaxaID	integer; identity	Unique integer identifier for each taxonomic classification	37
TSN	integer	Taxonomic Serial Number, from itis.gov	180542
Domain	nvarchar (50)	Scientific Domain name	*Eukarya*
Kingdom	nvarchar (50)	Scientific Kingdom name	*Animalia*
Phylum	nvarchar (50)	Scientific Phylum name	*Chordata*
Class	nvarchar (50)	Scientific Class name	*Mammalia*
Order	nvarchar (50)	Scientific Order name	*Carnivora*
Suborder	nvarchar (50)	Scientific Suborder name	*Caniformia*
Infraorder	nvarchar (50)	Scientific Infraorder name	n/a
Family	nvarchar (50)	Scientific Family name	*Ursidae*
Genus	nvarchar (50)	Scientific Genus name	*Ursus*
Species	nvarchar (100)	Scientific Species name	*Ursus maritimus*
Subspecies	nvarchar (150)	Scientific Subspecies name	n/a
CommonName	nvarchar (500)	Common name	Polar Bear
Synonyms	nvarchar (max)	Common synonyms	ours blanc
TaxaLink	nvarchar (500)	Hyperlink to the taxa report on itis.gov	www.itis.gov/...180542
TaxaComments	nvarchar (max)	Comments on the taxonomic classification	n/a

The extension of the CUAHSI data model necessitated corresponding changes to the CUAHSI ODM Data Loader and also a number of additions to the CUAHSI master list of Controlled Vocabulary (http://his.cuahsi.org/mastercvreg.html) – in all 133 additions and 17 edits to existing elements in the CUAHSI Observations Data Model controlled vocabulary tables were made to accommodate information from this investigation.

4.4.3 Chain-of-Custody Tracking

The importance of establishing data provenance is widely acknowledged, for quality control and quality assurance purposes, for questions of clarification and of collaboration, for discovering errors or making revisions, and for providing appropriate credit and citation for data use. As previously discussed, each data value stored in the COMIDA CAB project database is treated as an individual entity. As such, each data value has associated metadata describing the source of that value – who was the data collector/provider, what organization do they represent, and how can they be contacted. This type of provenance-tracking is in use in many data systems today, although admittedly not frequently enough, and might be considered the current best-practice.

The COMIDA CAB project team has taken the chain-of-custody approach one step further, however. Since this is a large project with diverse and complex project data, quality assurance and quality control assume increased importance. To aid project PIs in data validation and to allow for individual researchers to 'track' their input data as it moves through the data management workflow process, each data value maintains as metadata the name of the Excel file in which it was originally provided. Since the ultimate data archiving will include the 'raw' Excel files in addition to the project database, each data kernel may be traced back to its file of origin and to the PI who provided it. The Observations Data Model was amended slightly to explicitly assist in this chain-of-custody tracking effort: in the Source table, a mandatory attribute named SourceFile was added.

This enhanced chain-of-custody tracking affords the opportunity for each PI to review their data as it is represented within the complete project database. This data review was accomplished via a "data check" spreadsheet, prepared for and supplied to each PI populated with their data. The data check consists of 57 fields of data and metadata and was constructed as a view onto the project database with the data selected for inclusion via a query by PI name. As such, PIs can be confident that their own data has been represented faithfully and accurately within the database. Similarly, queries may be performed on the data by individual PI and/or by Excel file of origin.

4.5 COMMUNICATING RESULTS

4.5.1 Web-Based Data Access

The COMIDA CAB project is federally-funded; this means that the United States taxpayer owns the project results and data. Access to the complete project results is beneficial to the multiple project PIs, to other scientists, to regulators, and to stakeholders in the Chukchi Sea and its environs. As such, a project website was established as the primary project outreach platform, http://www.comidacab.org (Figure 30).

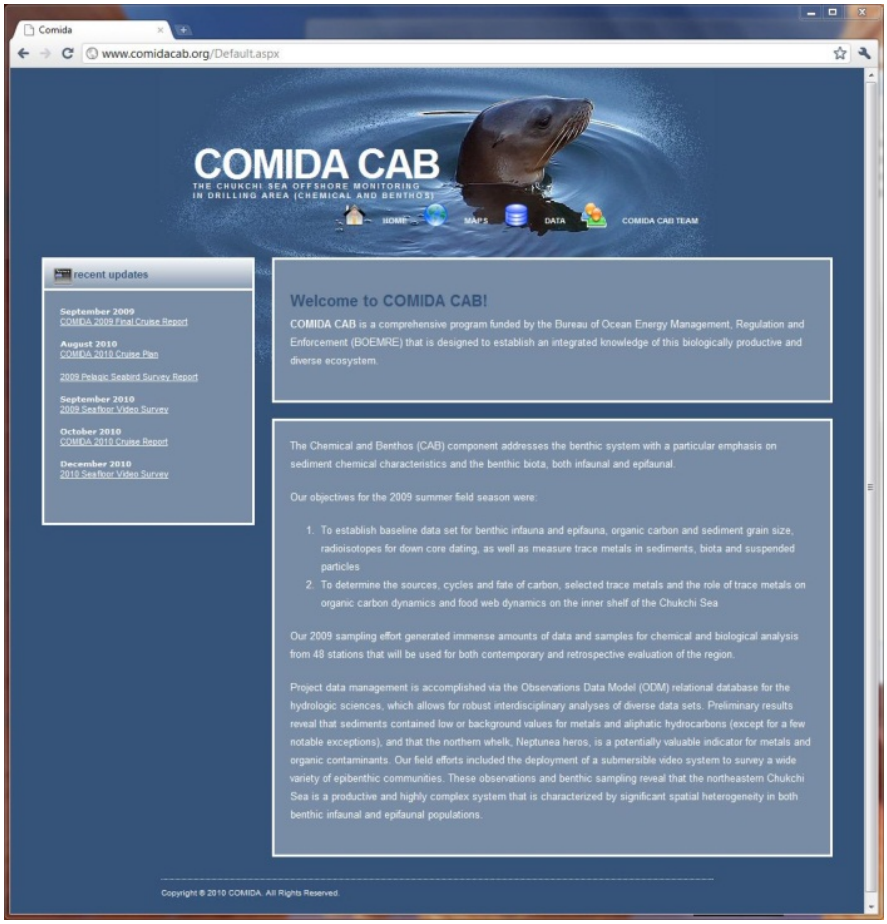


Figure 30. The COMIDA CAB project homepage, http://www.comidacab.org.

The website includes tabs for: (1) Home – providing a brief project overview plus recent updates; (2) Maps – linked to the ArcGIS Online community for sharing geographic data; (3) Data – linked to the Integrated Rule-Oriented Data System (iRODS) on the Corral Server of the Texas Advanced Computing Center (TACC); (4) Documents – to share project reports, posters and presentations; and (5) About – the COMIDA CAB project team. The project leverages ArcGIS Online (http://www.arcgis.com/), an online community for sharing geographic data, and iRODS, a grid software system for managing data on the web. iRODS is housed on the Corral Server of the Texas

Advanced Computing Center (TACC), a 1.2-petabyte set of disk arrays (https://goodnight.corral.tacc.utexas.edu/tacc/home/comidacab) (Figure 31).

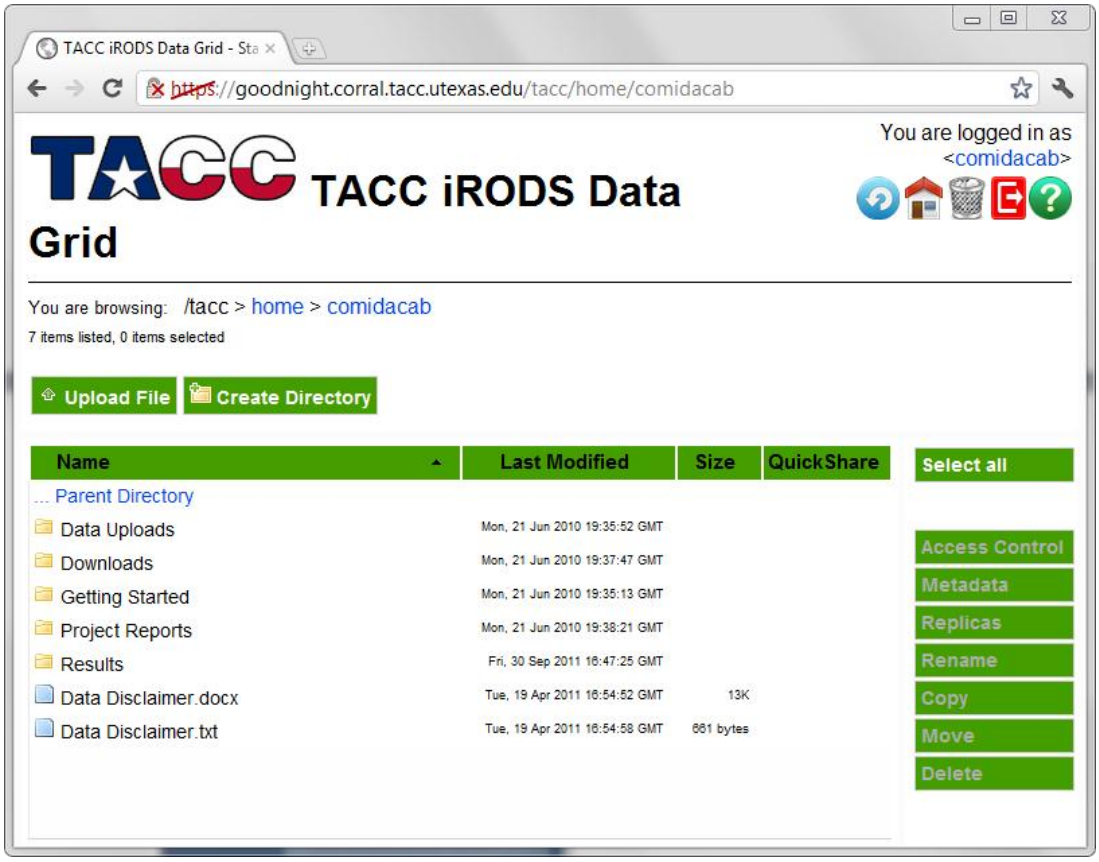


Figure 31. The COMIDA CAB iRODS online data storage system.

4.5.2 Data Visualization

ESRI ArcGIS 10 and the Geostatistical Analyst extension are being used for the analysis and visualization of observational data. The GIS provides for management, analysis, and display of spatially-referenced point samples and the interpolation of raster surfaces; these maps are useful for viewing and analyzing the data in a geospatial context

and also have significant value for communicating results to both technical and lay audiences. A selection of example visual representations is presented in Figure 32, Figure 33, and Figure 34.

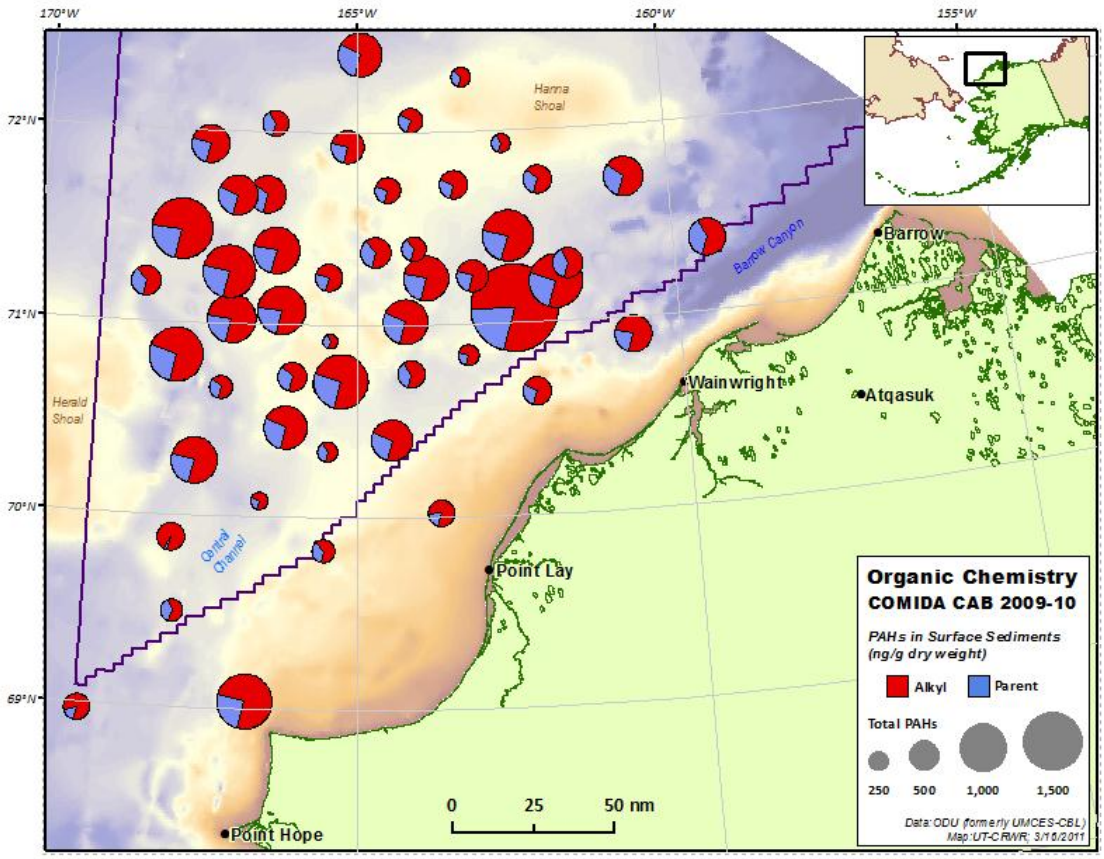


Figure 32. Examples visual representations of geographic data: (a) Polycyclic Aromatic Hydrocarbons in surface sediments.

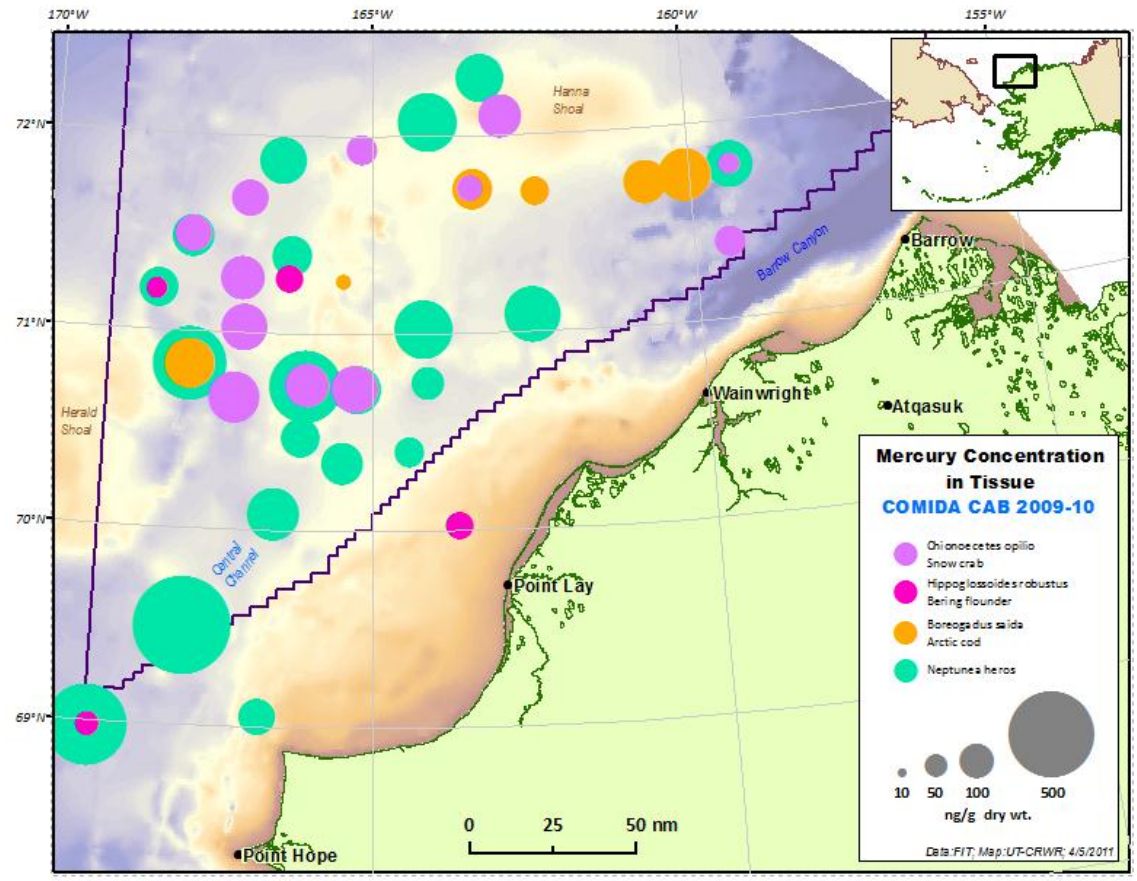


Figure 33. Examples visual representations of geographic data; (b) mercury concentration in organismal tissue.

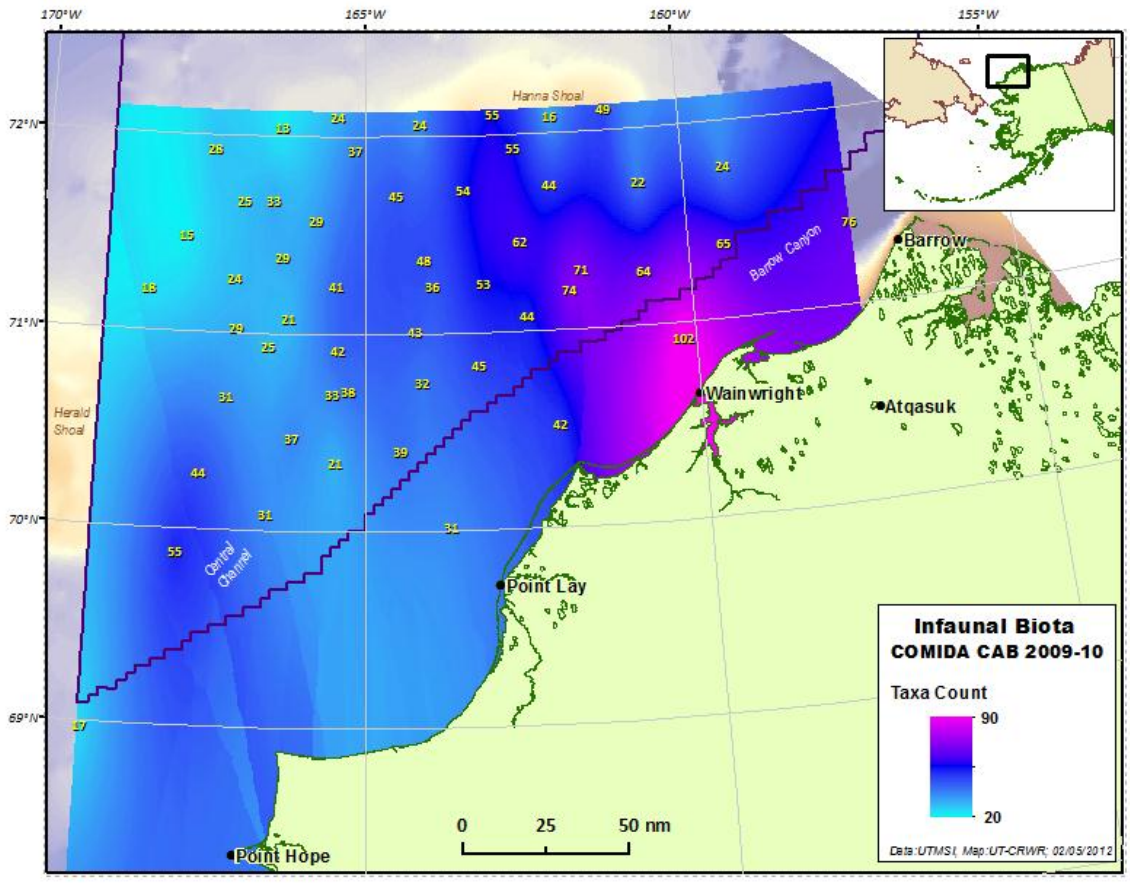


Figure 34. Examples visual representations of geographic data: (c) infaunal biota taxa count.

4.5.3 Data Archiving

In addition to these project-specific outreach efforts, the project data have been archived externally at the National Oceanographic Data Center (NODC), which has as its mission the provision of scientific stewardship of marine data and information and represents the world's largest holding of publicly-accessible oceanographic data. As

such, NODC serves as the national repository for information specific to the oceanographic discipline (NODC 2011).

The NODC operates as part of the N **NetCDF** ceanographic and Atmospheric Administration (NOAA). NOAA, and thus NODC, commonly employs the Network Common Data Format (NetCDF) to manage data. Featuring a binary structure well-suited for multidimensional data, NetCDF is an open-source data format that is widely-used in the atmospheric science community. It was developed by Unidata (2012) and has been adopted as a standard by the Open Geospatial Consortium (OGC 2012a). To assist oceanographers in submitting data to its data archive, the NODC has developed data submission templates which conform to Unidata's NetCDF Attribute Convention for Dataset Discovery (ACDD) and NetCDF Climate and Forecast (CF) conventions (NODC 2012). While not mandatory for data submission to the NODC, these NetCDF templates represent the current best practice for open-source and open-standard data sharing and data access.

Based on discussions with NODC data officials, the COMIDA CAB project team will submit a data package to the NODC upon approval for public data release. The package will consist of: (1) original PI data files as Microsoft Excel; (2) the project database in NetCDF format using the NODC data submission template; and (3) the final project report.

It is believed that the COMIDA CAB data submission package represents the most complete description available of the project results and affords the greatest flexibility for others to find, access, and ultimately use the project data from the NODC

archive. It is hoped that this novel submission package template will serve as a model for others submitting data to NODC to better leverage this valuable public resource for understanding and protecting our oceans.

The flow of information for the COMIDA CAB project can be envisioned as going from the bottom of the ocean…to a pile of Excel files…to a unified, homogenized project database…in a common data format…to a stable and persistent federal archive…to authoritatively establish the baseline conditions of the Chukchi Sea (Figure 35).

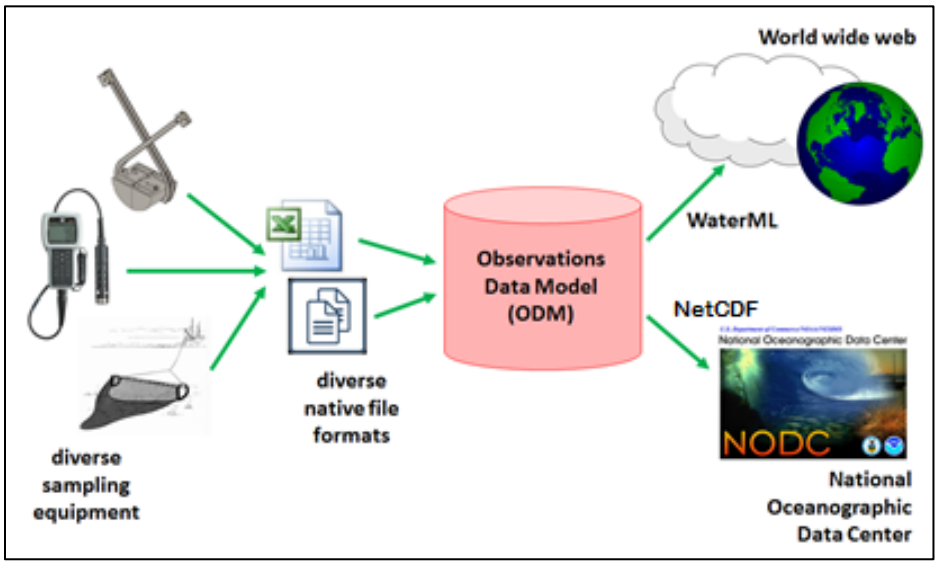


Figure 35. Flow of information for the COMIDA CAB project, from collection through homogenization and database creation to publishing and archiving.

4.6 CONCLUSION

Scientific oceanographic cruises have taken place for over 100 years and have yielded valuable insight into the patterns and processes of our planet's oceans. The value of this information is as high now as it has ever been due to multiple stressors on and numerous competing demands for oceanic resources. As such, thoughtful oceanographic data management is of critical importance. Presented herein is a workflow for observing the ocean environment, organizing and storing those observations, and communicating the resulting data and knowledge. Central to this workflow is the process of converting a heterogeneous "stack" of data files into a unified project database. The benefits of this process can be multiple: more transparent information, better decision support, better re-use opportunities, easier archiving, better documentation and metadata, better quality control and higher data quality.

On a more local level, thoughtful data management is of equal or greater importance for the COMIDA CAB project. One stated goal of the project is to establish quantitative baseline conditions of the Chukchi Sea ecosystem – of the epifaunal and infaunal biota, the sediment, and the water column – in advance of any potential oil and gas exploration and production in the region. As such, the data itself resulting from the COMIDA CAB cruises is a valuable product, additional analyses notwithstanding. By working with and providing data to the National Oceanographic Data Center, the stable long-term storage of the COMIDA CAB data is ensured, and the availability of these data

is furthered by embracing open-source data access and data standards in the manner described here.

One of the main reasons why the COMIDA CAB team was successful in securing the project grant was that they had a comprehensive idea about data management for the project. There was a significant question whether an observations model developed for CUAHSI HIS could be applied to the very different COMIDA CAB data. This investigation showed that, with the additions to handle taxonomy and chain of custody, the CUAHSI Observations Data Model proved to be quite robust and a modified version of the "star schema" of the ODM with the additional descriptor of a TaxaID proved to be a robust model for storing the project data even though 93% of them were physical CTD-type data and 7% were the much more complex biological kind of data.

The objectives of this paper were to present an approach to making observations of the ocean environment, to put forth a methodology for organizing and storing these observations, and to offer various avenues for communicating scientific results widely via the use of open standards. What was accomplished via this case study was the adaptation of the CUAHSI Observations Data Model for application with physical, chemical, and biological oceanographic data – a new extension of the CUAHSI Hydrologic Information System. Furthermore, the need for accommodating biological observations of the ocean environment in a cohesive project database drove further refinement of the BioODM data model, which was also amended to include better source tracking for the novel chain-of-custody approach introduced here. In this sense, a complete, real-world implementation

of an expanded Hydrologic Information System is presented, inclusive of biological and oceanographic data, for a multidisciplinary academic study.

Chapter 5: Managing Environmental Flows Information for Texas

5.1. Texas Environmental Flows Information System Case Study

Stakeholders and regulators across Texas are in the midst of a legislatively-driven process to determine the environmental flow needs of the bays, basins, and rivers of the state. Environmental flows are defined as "the quantity, timing, and quality of water flows required to sustain freshwater and estuarine ecosystems and the human livelihoods and well-being that depend on these ecosystems." (Brisbane Declaration 2007) As is common elsewhere, the environmental flow program in Texas includes analyses of hydrology and hydraulics, geomorphology and physical processes, water quality, biology, and the connectivity between and among these four primary disciplines. The synthesis of sometimes disparate findings from these disciplines stands to be one of the most challenging and most important steps of developing instream flow recommendations.

Given the large spatial and temporal scales of analysis necessary for sufficiently detailed study of environmental flow issues, a relative paucity of data exists to support these analyses. This challenge is acutely evident in the determination of flow-biota linkages and the assessment of habitat availability and suitability. As such, there is a need for the development of tools and systems to organize, share, and synthesize information. The case study presented here is an effort to have biological information for

environmental flow studies organized in a manner consistent with that currently used for physical and chemical information.

The Environmental Flows Information System for Texas project seeks to provide improved data access and integration to aid stakeholder committees, expert science teams, and the Texas Commission on Environmental Quality in their collective efforts to determine statewide environmental flow needs. In addition, a demonstration project for the river basin and bay system consisting of the Trinity and San Jacinto Rivers and Galveston Bay was conducted which sought to organize and foster access to documents, reports, and studies.

Continuing the V-shaped exposition, this chapter branches out further to a case study which considers observations data for aquatic biology alongside other types of information and which serves a wider audience of stakeholders and practitioners in the field of environmental flows.

5.2. Example Use Cases for Environmental Flows

Often when designing software or information systems, a useful first step is to determine the expected typical usage. This determination can be accomplished via the development of typical usage scenarios: who will use the system and what do they hope to do with it? What is the typical user's skill level? Do they want to find something, share something, analyze something? This process is called "defining a use case." (Jacobson et al 1992) Although use case formats vary, common elements include:

- Usage description (goal)
- Actors (user profile)
- Assumptions
- Workflow (steps to achieve the goal)
- Variations and Special Requirements.

The following are some example use cases: situations from the perspective of a user (scientist, researcher, stakeholder, general public, etc) making a query on the Texas Environmental Flows Information System (Table 11). These use cases may also be found online at: http://www.cuahsi.org/docs/EnvironmentalFlows.pdf

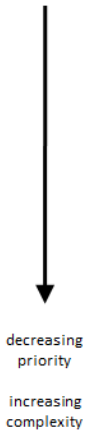
Table 11. Example queries for the study of environmental flows.

Example Query	General Query Type
What data were collected on the Lower Sabine River study?	All variables for all time for all space within a specified geographic extent (basin, reach, river, site)
I'm interested in fish distribution; what largemouth bass data are available in Texas?	One variable for all time for all space.
What sites measure channel bed substrate?	Sites (space), identified by variable.
What invertebrate data is available in Texas? How about mesohabitat data? Sunfish family?	All data by variable group (fish taxonomy, habitat).
Was a report or article written based on the data at this site?	Specific document by site and/or sample.
What reports or articles are available for the Trinity River basin?	All documents by specified geographic extent (basin, etc).
For Guadalupe bass, what is the trophic group (omnivore, piscivore, herbivore etc); conservation status (threatened, endangered, etc); mesohabitat guild (fast riffles, deep pools, etc); habitat suitability criteria (acceptable depth, velocity, and substrate, etc)?	Data value (variable) linked to supporting relational database.

Another use case developed for the Texas Environmental Flows Information System concerns coastal processes, relevant to the determination of environmental flow needs in Texas since the "instream" water is delivered to coastal bays and estuaries as "freshwater inflows" and thus acts as a significant control on salinity levels (Table 12). The previous use cases discussed are framed around typical queries that a scientist might be seeking data in order to answer. In the following use cases, these typical queries are

still present in the Example column, but have been structured to demonstrate sample queries which may draw from a range of data services, sampling sites, and variables.

Table 12. Example use case displaying cardinality of data services, sites, and variables.



	Service	Site	Variable	Example
1	1	1	1	Salinity data for Galveston Pier 21 from TCOON
2	1	1	many	Salinity, wind speed, and water temperature data for Galveston Pier 21 from TCOON
3	1	many	1	Salinity data for all Galveston Bay sites from TCOON
4	1	many	many	Salinity, wind speed, and water temperature data for all Galveston Bay sites from TCOON
5	many	1	1	Salinity data for Galveston Pier 21 from TCOON, TPWD, and TWDB
6	many	1	many	Salinity, wind speed, and water temperature data for Galveston Pier 21 from TCOON, TPWD, and TWDB
7	many	many	1	Salinity data for all Galveston Bay sites from TCOON, TPWD, and TWDB
8	many	many	many	Salinity, wind speed, and water temperature data for all Galveston Bay sites from TCOON, TPWD, and TWDB

These use cases demonstrate one distinct advantage of a database and an information system over individual data files – the ability to access all data pertinent to a specific question (e.g.: “What is the salinity in Galveston Bay?”) regardless of whether the data come from one data source or multiple sources. Table 12 also shows the cardinality between entities in the database – the numeric relationship between each attribute. For example, Use Case #1 can be accomplished using data from only one service (Texas Coastal Ocean Observation Network, TCOON), at one site (Galveston

Pier 21), for one variable (salinity), whereas Use Case #8 requires data from many services (TCOON, Texas Parks and Wildlife Department (TPWD), and Texas Water Development Board (TWDB)), for many sites (all locations within the extent of Galveston Bay), for many variables (salinity, wind speed, and water temperature). In this regard, Use Case #1 displays a one-to-one cardinality between site and variable, Use Case #8 displays a many-to-many cardinality between site and variable, and the intermediate Use Case #6 displays a one-to-many cardinality between site and variable.

5.3. LOWER SABINE RIVER CASE STUDY

5.3.1 Background and Purpose

The discussion that follows presents a case study from the Texas Instream Flows Program for the Lower Sabine River in Texas and Louisiana. The data were collected by staff from the Sabine River Authority of Texas, the Texas Parks and Wildlife Department, TCEQ, and TWDB and the analysis described here was conducted by the author.

Of the first two Bay and Basin Expert Science Teams (BBESTs) to deliver reports of their findings to TCEQ regarding environmental flow recommendations, one (the Trinity/San Jacinto/Galveston BBEST) was unable to reach a consensus with their recommendations. This was attributed partly to a fine-scale analysis of hydrology with little or no connection to the aquatic biology of the basin and bay system. One benefit of the Environmental Flows Information System is to aid in the synthesis and analysis of the

available biological datasets in Texas, incorporating data from state agencies all the way down to individual academic researchers.

An example biological data analysis was performed using data collected from the Texas Instream Flow Program on the Lower Sabine River in 2006. "The study on the Lower Sabine River was prioritized based on the potential for water transfers within the Sabine Basin, proposed inter-basin water transfer projects, and Federal Energy Regulatory Commission hydropower relicensing of the Toledo Bend Dam." (SRATX 2007). These data were obtained by request from the TWDB and the purpose of presenting this limited analysis here is to showcase one example of the type of basin understanding which could be achieved via improved access to the state's aquatic biology data. These previously-inaccessible data have now been made publicly-accessible via EFIS (http://efis.crwr.utexas.edu/downloads/SanAntonio_baseline_fish_sampling.zip), via the data.crwr website (http://data.crwr.utexas.edu/tifp_sabine.html), and via CUAHSI HIS Central (http://hiscentral.cuahsi.org/pub_network.aspx?n=50).

The purpose of this case study is thus two-fold: (1) to provide a detailed view of the types of information used, the types of questions posed, and the types of analyses conducted to aid in the determination of environmental flow needs; and (2) to demonstrate how an improved Hydrologic Information System which can accommodate biological data can be of use in these efforts. Of particular importance is the inclusion of both taxonomy and traits – the genus and species of all fish observed is recorded along with the organism count and the minimum and maximum total lengths of each fish species. In this configuration, taxonomy exerts primacy over trait in the sense that a

researcher is much more likely to perform a query on the database by species ("What was the maximum length of a harlequin darter in study reach 5020?") than by trait ("Of all fishes observed over 30 centimeters in length, what percent of them were longnose gar?").

5.3.2 Lower Sabine River Instream Flow Study Observations

Over 8 field days, 165 samples were collected at 8 study reaches with 147 of those samples (89%) yielding fish (Figure 36, Figure 37). Fish were collected using seine nets and backpack- and boat-mounted electroshock units.

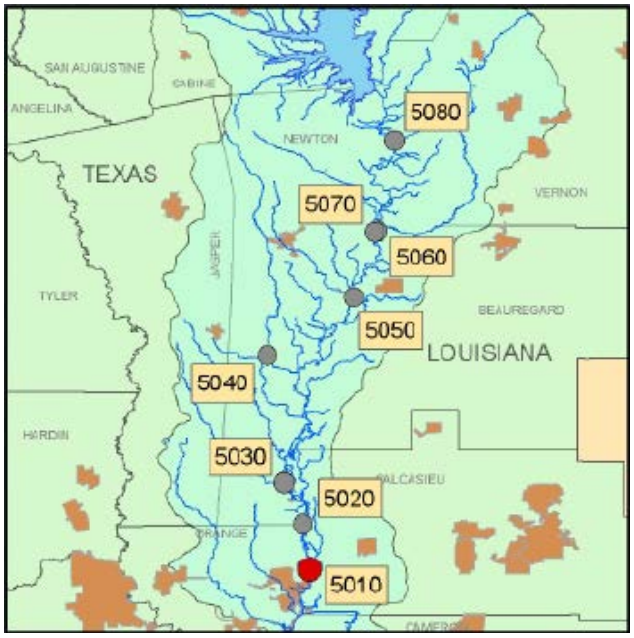


Figure 36. Lower Sabine River baseline fish sampling sites, May to September 2006 (SRATX 2007). At the southern end of the map is Sabine Lake salt water estuary and at the northern end is the Toledo Bend Reservoir, formed by the Toledo Bend hydropower dam.

Figure 37. Sabine River sampling locations. Noteworthy in this image is the referential integrity of the handheld GPS unit used to identify the location of samples marked with identifiers such as “right bank,” “left sand bar,” and “center of tributary channel.” (SRATX 2007)

In addition to fishes, physical habitat data were collected on mesohabitat type (pool, backwater, run, or riffle), on water depth, on stream velocity (at either 60% or 20% and 80% of water depth based on the depth encountered), on channel substrate material (silt/clay, sand, gravel, rubble/cobble, boulder, or bedrock), on cover type (such as overhanging vegetation, undercut banks, submerged vegetation, submerged rocks and logs, and floating debris), and embededness (the degree to which larger substrate grains are covered by fine sediment) (Table 15). Furthermore, data were collected on the level of sampling effort – seine length (the cross-stream width dimension of the seine net), haul length (the distance the net was dragged), and shock distance (the length along which electrical current is being applied in the water).

The types of data collected in this study on fish and habitat are representative of those commonly collected across the globe for environmental flow assessments. Although conditions vary from study-to-study and site-to-site, a few common categories of research questions are asked and answered using data such as these (Table 13). This list is by no means exhaustive, and some of these questions are addressed somewhat is the discussion which ensues.

Table 13. Example common research questions posed in an environmental flows assessment and the corresponding data requirements.

Research Question	**Data Requirements**
What is the (fish, mussel, invertebrate, etc) community structure?	Species counts and richness, diversity indices
What are the habitat drivers and controls?	Stream and channel habitat data
What is the distribution and relative abundance of native/ invasive species?	Native/invasive status
What is the status of species of conservation concern?	State and federal lists of threatened and endangered species
What is the long-term habitat quality?	Tolerance/intolerance thresholds, habitat suitability indices

5.3.3 Fish Community Analysis and Characterization

A total of 5831 fish were observed representing 58 species (averaging 40 fish per sample). On average, 729±433 fish were collected per study reach with a range of 72 to 1365 fish. The average study reach had 24±8 species represented with a range of 10 to

36 species present (out of 58 total across all sites). The five most abundant species collectively accounted for nearly 70% of the fishes observed (Table 14, Figure 38).

Figure 38. (a) Blacktail shiner (*Cyprinella venusta*); (b) Bullhead minnow (*Pimephales vigilax*); (c) Bay anchovy(*Anchoa mitchilli*); (d) Spotted bass (*Micropterus punctulatus*); and (e) Sabine shiner (*Notropis sabinae*). Not to scale. Figures a, b, d, e from (Thomas et al. 2007); Figure c from (Wood and Williams 2005).

Table 14. Fish species abundance in the Lower Sabine River.

Species	Common Name	Figure	Family	Count	Relative Abundance
Cyprinella venusta	blacktail shiner	a	*Cyprinidae*	2101	36.0%
Pimephales vigilax	bullhead minnow	b	*Cyprinidae*	595	10.2%
Anchoa mitchilli	bay anchovy	c	*Engraulidae*	542	9.3%
Micropterus punctulatus	spotted bass	d	*Centrachidae*	446	7.6%
Notropis sabinae	Sabine shiner	e	*Cyprinidae*	365	6.3%
All others				1782	30.6%
Total				**5831**	**100.0%**

Table 15. Variables in the Lower Sabine River observations database.

Biological	Physical	Sampling Effort
scientific name (genus and species)	water depth	seine length
common name	velocity at 20% depth	haul length
count	velocity at 60% depth	shock distance
minimum total length	velocity at 80% depth	shock time
maximum total length	substrate type	
	habitat type	
	cover type	
	embededness	

The fish community structure in each of the eight study reaches was examined using three methods: the Shannon-Weiner Diversity Index (H'), Pielou's Evenness Index (J'), and the Simpson Dominance Index (D). Calculation of these indices relies on the total organism Count (N) and the total Species Richness (S) (Table 16). The Shannon-Weiner Diversity Index was originally developed to quantify entropy in written text, whereby the uncertainty associated with correctly predicting the next letter in a string is greater if more different letters are present (Shannon 1948). In this sense, the index itself increases as the number of letters (or species) increases. Pielou's Evenness Index is a measure of how equal the relative abundance is of various species within a community (Pielou 1966). If the same number of every species is present, the community is completely even and therefore J'=1.0, whereas if one species dominates, J' approaches zero. Similarly, a value of D=1.0 in the Simpson Dominance Index represents infinite species diversity whereas zero represents no diversity whatsoever (Simpson 1949).

A limited investigation in possible spatial patterns of the fish community structure yielded the following results. The downstream-most study reach 5010, closest to the Sabine Lake salt water estuary, exhibited high fish presence (count), moderate species richness, and low diversity. This reach was dominated by bay anchovies, a coastal species tolerant of a wide range of salinities. The next three reaches moving upstream (5020 through 5040) exhibited higher fish presence but lower species richness. Cyprinids, especially blacktail shiner, dominated these reaches and thus caused lower diversity, lower evenness, and more dominance. The next three reaches (5050 through 5070), exhibited higher fish counts and lower diversity with high relative abundance of

blacktail shiner, bullhead minnow, and Sabine shiner. The upstream-most study reach 5080, located less than 20 miles below the Toledo Bend hydropower dam, exhibited a very low fish count and a corresponding very low species richness. This reach was dominated by blacktail shiner and longear sunfish.

Table 16. Lower Sabine River fish community characterization.

Study Reach	Count	Species Richness	Shannon-Weiner Diversity Index	Pielou's Evenness Index	Simpson Dominance Index
	N	S	H'	J'	D
5010	1096	21	1.700	0.558	0.707
5020	545	30	2.454	0.721	0.869
5030	401	28	2.342	0.703	0.841
5040	527	25	2.405	0.747	0.840
5050	1130	36	1.572	0.439	0.557
5060	695	17	1.166	0.412	0.553
5070	1365	24	1.706	0.537	0.692
5080	72	10	1.566	0.680	0.686

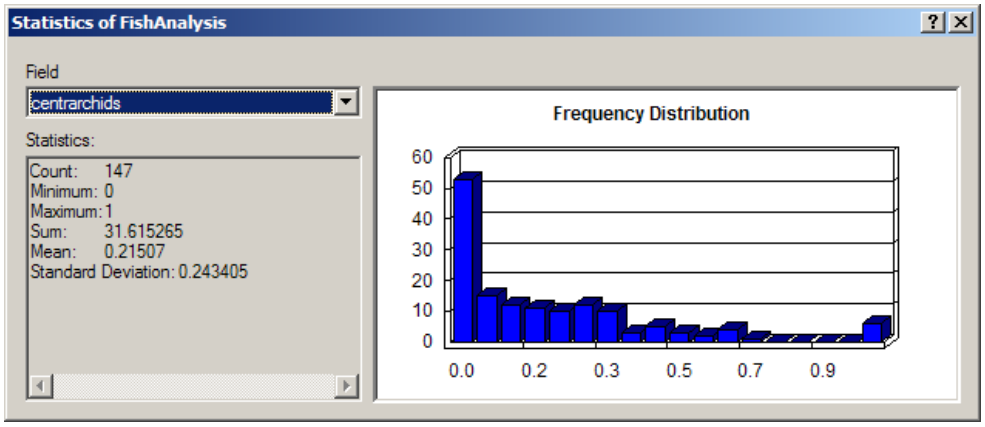


Figure 39. Frequency distribution of the relative abundance of the family *Centrarchidae* (sunfishes and bass) in the Lower Sabine River study area.

Across all sites, 889 fishes of the family *Centrarchidae* (sunfishes and bass) were observed with a relative abundance of 22% ±24% (Figure 39). Additionally, three blue suckers (*Cycleptus elongatus*) were observed, a state-listed threatened species. The presence and geographic distribution of imperiled species (i.e., those species classified as vulnerable, threatened, or endangered) is often an important factor in the determination of environmental flow regimes and in the prioritization of stream restoration and conservation efforts. However, there is some disagreement regarding the public release of observations data for imperiled organisms which specifies a known location for those organisms; this concern stems from a desire to protect the imperiled organisms from disturbance from curious and interested parties.

The only fish non-native to the Sabine River Basin observed was *Menidia beryllina* (inland silverside); 192 of these individuals were collected ranging from 0 to 90% of the sample population with a mean of 3% ± 12% (Figure 40). *M. beryllina* was

"originally found in coastal waters and upstream in coastal streams along the Atlantic and Gulf coasts; widely introduced into freshwater impoundments." (Hubbs et al. 1991; Thomas et al. 2007)

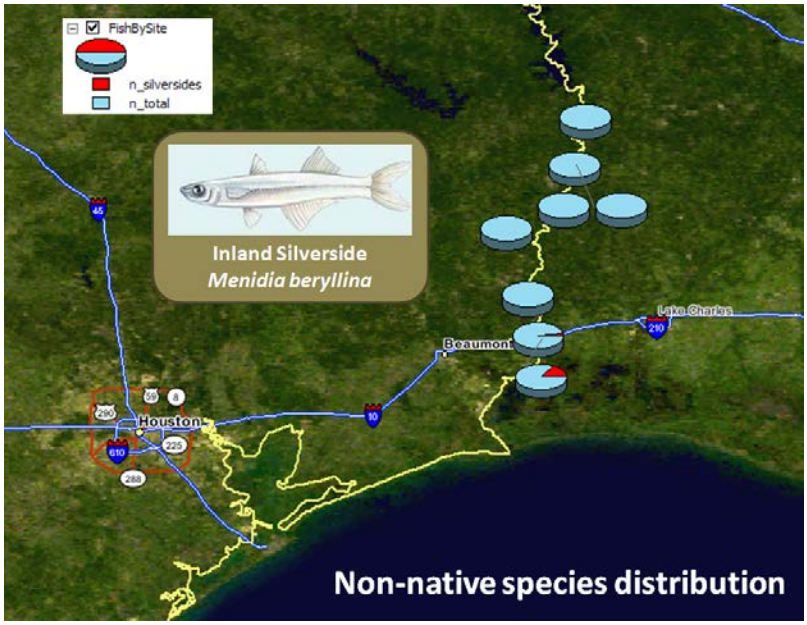


Figure 40. Distribution of the non-native inland silverside (*Menidia beryllina*) fish species in the Lower Sabine River Basin, Texas/Louisiana.

From an environmental flows perspective, two broad conclusions may be drawn from the analyses presented here. First, the Toledo Bend hydropower dam appears to be having an impact on downstream fish abundance and species richness, so a consideration of the release schedule and potential dam reoperation merits attention. Second, the inland silverside has had some, but limited, success in invasion in the Lower Sabine River.

Thus, an examination of freshwater inflows to Sabine Lake is warranted, as is an investigation into the salinity balance and its effect on invasive species control.

5.3.4 Linking Observations Data to Maps and Documents

"The Texas Environmental Flows Information System and a corresponding Texas Water Development Board-sponsored project to develop a Texas Hydrologic Information System both seek to organize and facilitate spatially-explicit access to water data. One common means of accessing these data is through a map interface. It is sometimes the case that the data collected and made available through such information systems were aggregated and analyzed into a journal article, thesis, research report, data summary, study, or other similar document, and the analyses, conclusions, and recommendations from these documents often provide added value. Thus, it is worthwhile to provide parallel access to both the data and the knowledge products derived from that data.

"A prototype linkage to a georeferenced digital archive of documents (orange polygon in Figure 41) was developed using the same map interface that hosts the data so the user can access both types of information concurrently. The documents have been represented by polygons instead of points as polygons are believed to be more spatially-explicit and thus provide a more accurate geographic representation of the study area addressed in any given document.

Figure 41. KML-based polygonal geographic representation of the Lower Sabine River Instream Flow Study, depicted alongside the study sampling sites; linkages to both the data and the document are provided from the map interface. (Hersh et al. 2008)

"Thus far, example geographic representation of digital documents has been provided via KML (formerly, Keyhole Markup Language) and ESRI shapefiles; the use of Web Feature Services (WFS) is currently being explored by CRWR and the Texas Natural Resources Information System (TNRIS). KML is a geographic language developed and made popular by Google in their Google Earth software; KML has recently been accepted as an Open Geospatial Consortium (OGC) standard (OGC 2008). Shapefiles are a proprietary vector data format of the ESRI software company, providers of the popular ArcGIS line (ESRI 1998). WFS is an OGC interface standard for the communication of geographic data designed to support interoperability in that it is not tied to any specific software program, operating system, or platform (OGC 2008b)." (Hersh et al. 2008)

5.4. THE TEXAS ENVIRONMENTAL FLOWS INFORMATION SYSTEM

The current Senate Bill 3 process has had mixed success in establishing environmental flow needs in Texas bay and basin systems. In a memo to the stakeholder committees and expert science teams, The SB3 Science Advisory Committee has noted that "…issues have arisen with regard to the lack of sufficient site-specific scientific data and analyses describing the essential relationships between environmental flows and the actual needs of aquatic organisms in those systems." (SAC 2010) One step toward solving this problem is to bring biological data into a structured format to stand on the same footing as hydrologic data.

A prototype environmental flows information system is developed for the State of Texas that incorporates relevant known available datasets from federal, state, academic, river basin, and local sources (http://efis.crwr.utexas.edu) (Figure 42). Tools are developed to assist in the publishing, visualization, and access of data and documents via map-based, spreadsheet-based, and other methods. This project puts forth the concept of a Water Information System comprised of three components: (1) Geographic Information Systems (GIS) for geographic data; (2) Hydrologic Information Systems (HIS) for observations data; and (3) Digital Libraries for digital assets (documents, images, videos).

Six information types are included in the Information System:

1. Point observations data (communicated via the WaterML web language and stored in the CUAHSI Observations Data Model) (CUAHSI 2009);

2. Geographic data (such as shapefiles, feature classes, KML, WFS/WCS/WMS);
3. Documents (stored in the DSpace digital archive);
4. Tables (such as fishes conservation status and trophic guilds);
5. Tools (such as a Microsoft Excel-based Calculator for Low Flows); and
6. Links (including the Fishes of Texas project, the Indicators of Hydrologic Alteration model, and many others).

Altogether, the Information System contains nearly 100 components from over 25 contributors, including: state sources (e.g.: Texas Commission on Environmental Quality, Texas Water Development Board, Texas Parks and Wildlife Department, Texas Coastal Ocean Observation Network, and Texas Natural Resource Information System); federal sources (e.g.: United States Geological Survey, US Environmental Protection Agency, National Weather Service, National Oceanographic and Atmospheric Administration, and US Fish and Wildlife Service); academic (e.g.: University of Texas, Texas A&M University, Texas State University, University of New Orleans, and CUAHSI); non-governmental organization sources (e.g.: World Wildlife Fund and The Nature Conservancy); and river authorities (e.g.: San Antonio River Authority, Sabine River Authority of Texas. This content may be accessed through one of four avenues: (1) Web page; (2) Interactive Map Viewer; (3) Digital Library; and (4) HydroPortal.

The tables in EFIS provide an additional level of analytical support not typically found in present-day Hydrologic Information Systems. A great deal of current thinking in biological research focuses on the structure and *function* of ecosystems, on trophic

redundancy, and on the ecosystem services offered by various species and organisms. In this sense, higher-level knowledge products such as compilations of trophic groups, mesohabitat guilds, reproductive guilds, conservation status, and functional feeding groups provide additional meaning and add value to the raw observations data. In this regard, the design and content of EFIS supports analysis on three tiers: (1) at the organism level (e.g., individual organism traits); (2) at the community level (e.g., abundance and diversity metrics); and (3) at the functional level (e.g., across guilds and trophic groups).

An Interactive Map Viewer was developed which incorporates hydrologic basemap data for the United States developed and hosted by ESRI, overlain by observations data developed and hosted by the Center for Research in Water Resources at the University of Texas at Austin (Figure 43). Geographic and observations data is also available via the HydroPortal, a customization of the ESRI Geoportal Toolkit extension (Figure 44). Finally, a Digital Repository was developed in conjunction with the University of Texas Library system based on the open-source DSpace digital archive system (Figure 45) (DSpace 2009).

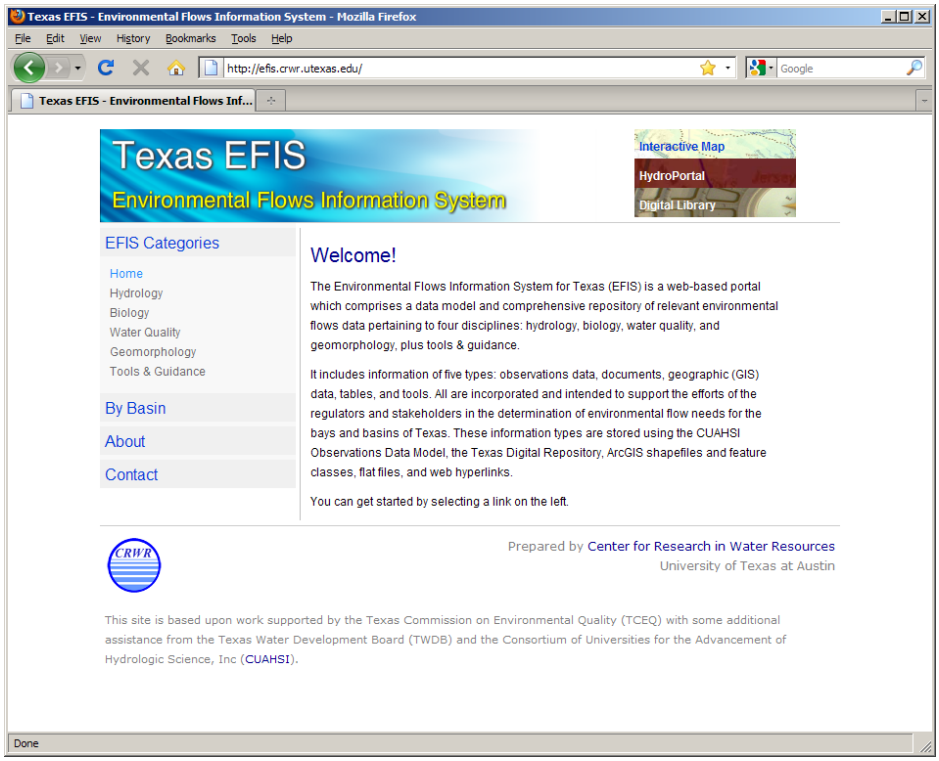


Figure 42. Environmental Flows Information System for Texas site homepage: http://efis.crwr.utexas.edu.

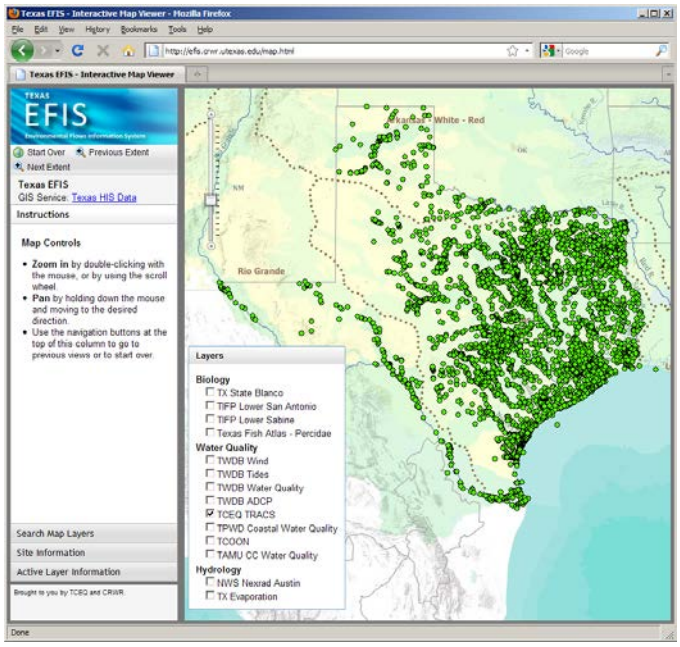


Figure 43. EFIS Interactive Map: http://efis.crwr.utexas.edu/map.html.

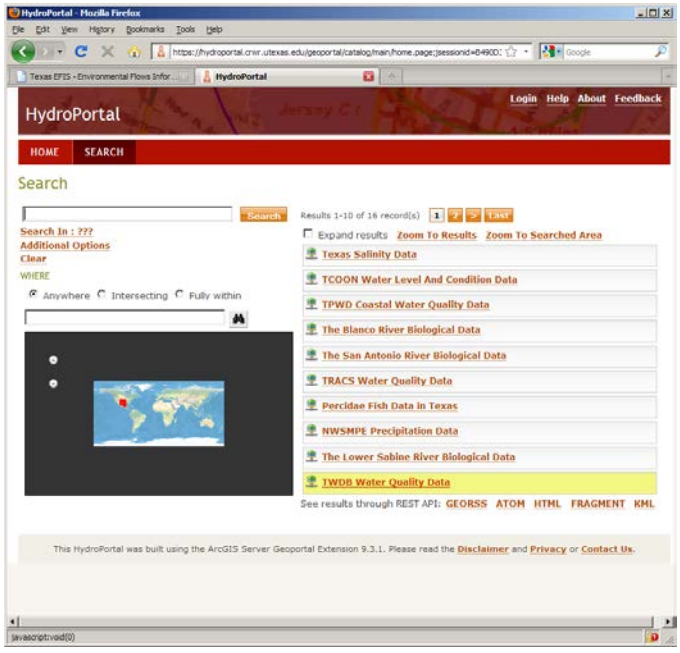


Figure 44. Environmental Flows Information System for Texas HydroPortal.

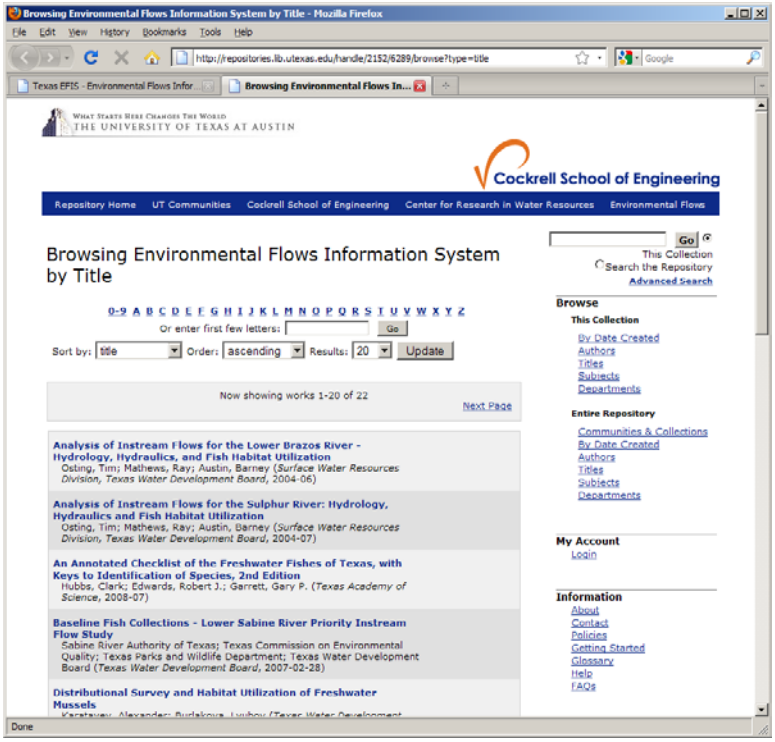


Figure 45. Environmental Flows Information System for Texas Digital Repository

5.5. THE CALCULATOR FOR LOW FLOWS

One contribution of the CUAHSI HIS project was the conceptualization and development of a Services-Oriented Architecture (SOA) for the communication of water data over the internet via a standard language, WaterML (Figure 46). The CUAHSI SOA includes HydroServer for data storage, HydroCatalog for data inventory and discovery, and HydroDesktop for data access and analysis. The significance of an SOA is that it enables data to be made accessible and communicated over the internet via web services, not just stored locally. In recognition of the value of SOA, the USGS has adopted

WaterML as its standard for online data sharing and is currently publishing and sharing data in its National Water Information System (NWIS) via WaterML.

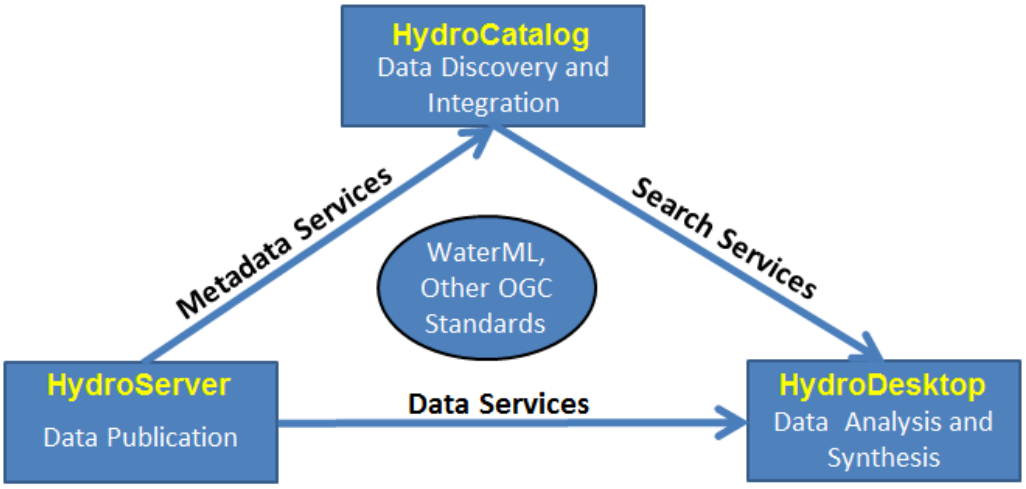


Figure 46. The CUAHSI HIS Services-Oriented Architecture (CUAHSI 2012).

As part of the EFIS project, a Microsoft Excel-based application called the Calculator for Low Flows was developed. CaLF is discussed here as an example of a relatively simple tool which leverages a significant federal data reserve via water web services to readily provide exactly the kind of information needed by environmental flow practitioners in their determination of environmental flow needs. CaLF is a tool for: (1) downloading USGS daily streamflow data; (2) calculating the seven-day two-year low flow (7Q2); (3) calculating and plotting the flow duration curve; (4) calculating the harmonic mean; (5) calculating the Lyons' method monthly minimum streamflow and the modified Lyons' method streamflow and adjusting them via a Drainage Area Ratio if

desired; and (6) graphing these two minimum flows. CaLF is designed in part to aid in the parameterization of the Hydrology-Based Environmental Flow Regime tool (HEFR) (SAC 2011).

The CaLF tool uses web services to download U.S. Geological Survey (USGS) mean daily streamflow data over an internet connection (Figure 47). This data is imported to the CaLF tool and manipulated through Visual Basic programming. CaLF is based on the technology of HydroObjects (Whiteaker 2008) and LDCurve, a tool for automatically creating bacterial load duration curves for water quality segments in the State of Texas. (Johnson 2009, Johnson et al. 2008) CaLF is accessible at: http://efis.crwr.utexas.edu/tools_guidance.html or at http://tools.crwr.utexas.edu/CaLF/.

The 7Q2 is calculated and reported on the "7Q2" worksheet by calculating the seven-day minimum flow within each year of record (the 7Q1), then determining the value in this new series with a 2-year return interval (i.e., the median of the 7Q1 values). The 7Q2 is defined as "The lowest average stream flow for seven consecutive days with a recurrence interval of two years, as statistically determined from historical data" and is relevant in Texas because "some water quality standards do not apply at stream flows which are less than the 7Q2 flow." (TCEQ 2000)

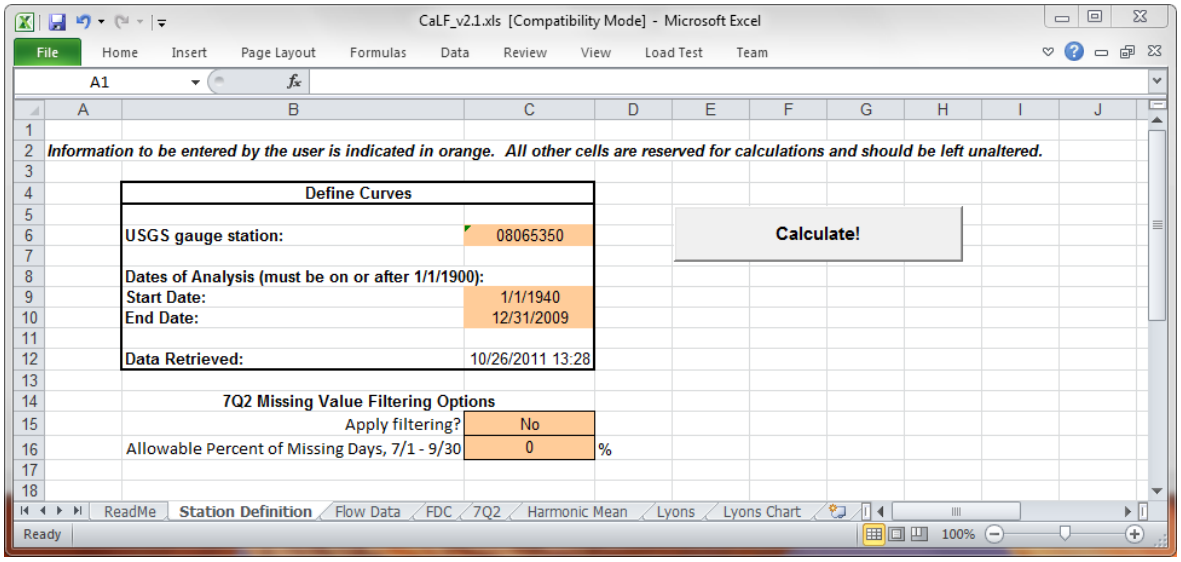


Figure 47. Station Definition tab of the Calculator for Low Flows tool.

The Lyons' flows are calculated as 40% of the monthly median streamflow for October through February and 60% of the monthly median streamflow for March through September. The Modified Lyons' method simply replaces any monthly calculated Lyons' flows which fall below the 7Q2 with the calculated 7Q2 value (effectively using the 7Q2 minimum flow as an absolute floor) (Figure 48). If desired, the Lyons' and Modified Lyons' flows can be adjusted via a user-input Drainage Area Ratio (DAR) on the "Lyons" worksheet. The DAR can be entered directly, or the Diversion Point Area and the Stream Gage Area can be entered and the DAR will be calculated accordingly.

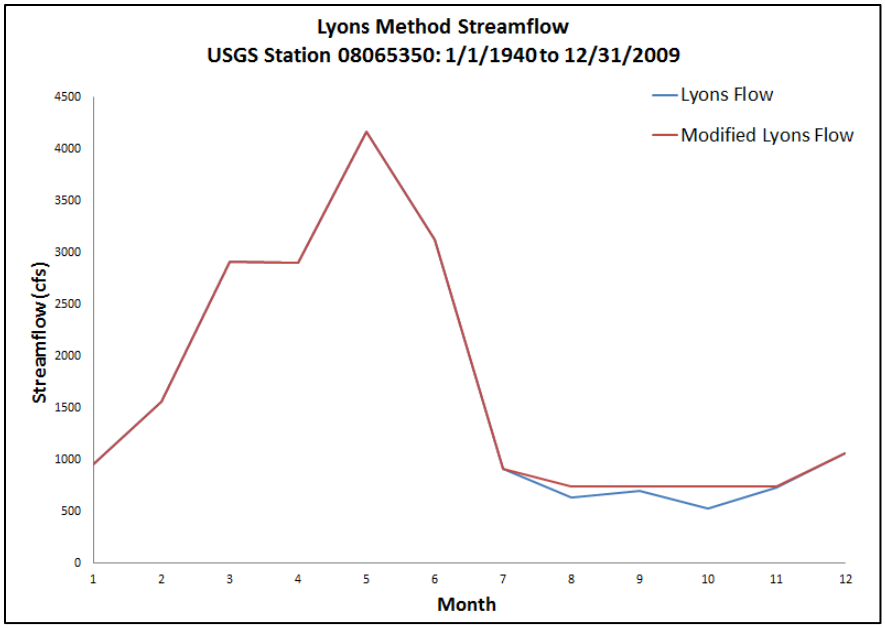


Figure 48. Example output for the Lyons method and modified Lyons method streamflows, USGS #08065350, Trinity Rv nr Crockett, TX. Note that the 7Q2 streamflow is used in the months of August, September, and October per the modified Lyons methodology.

Water web services are critical in the success of the CUAHSI Hydrologic Information System project. As has been shown via the CaLF example, water web services can play an important role in a wide range of disciplines – supporting analysis, modeling, and interpretation of data via the internet. While CaLF is a relatively simple tool designed to support the work of an environmental flows practitioner, there is no reason any number of similar tools could not be developed which use USGS streamflow data for any other particular application, nor any reason that a similar tool couldn't be developed to leverage any other water web service.

5.6 DIGITAL REPOSITORIES

Access to existing data and documents is a valuable and necessary tool for future scientific and engineering analyses. Some documentation is readily available through various means. Much is currently unavailable, however – a significant detriment to accomplishing the goal of establishing environmental flow needs. In addition to the Digital Repository created for the Environmental Flows Information System of Texas project, a demonstration project for the river basin and bay system consisting of the Trinity and San Jacinto Rivers and Galveston Bay was conducted which sought to organize and foster access to documents, reports, and studies.

The objectives of this demonstration project were to create a comprehensive Environmental Flows Document Model that would provide the format and organizing scheme for the incorporation of information from the multiple relevant disciplines; compile representative existing information on the hydrology, biology, physical habitat, physical processes (geomorphology), and chemical processes (water quality, aquatic life uses, etc.) of the study area; and deliver a prototype temporally- and spatially-explicit annotated bibliography of documents, reports, studies, and journal articles pertaining to the study area. In conjunction with the University of Texas Libraries, the DSpace digital repository system was used to capture, store, index, preserve, and redistribute documents.

The Trinity River Basin document collection is an early adopter of a much larger DSpace adoption effort at the University of Texas at Austin and in cooperation with the Texas Digital Library, “a multi-university consortium providing the digital infrastructure

to support an online scholarly community for higher education in Texas" (http://www.tdl.org) (Figure 49). The goals of the Texas Digital Library are very much aligned with the goals of the Trinity River demonstration project. TDL seeks to provide: (1) Access to a wide range of digital materials, (2) Long-term preservation of digital collections, (3) Support for the scholarly community, and (4) Aggregation of resources (TDL 2008).

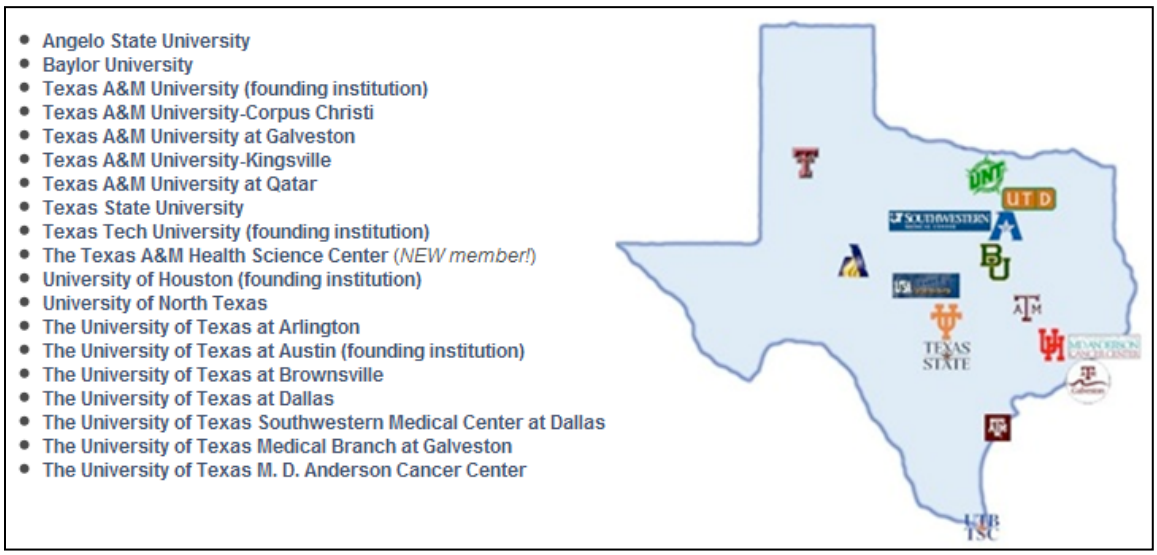


Figure 49. Texas Digital Library members.

The UT-Austin DSpace Repository (http://repositories.lib.utexas.edu/) was initiated on September 1, 2008; it's stated purpose is "to collect, record, provide access to, and archive the scholarly and research works of the University of Texas at Austin, as well as works that reflect the intellectual and service environment of the campus." (UT Libraries 2008). The Trinity River Basin project falls into the latter category.

Additional information is kept with the data file as metadata. The term metadata means "data about data" and is used to classify content for organization and retrieval. The Trinity River Basin prototype document management system metadata is as follows:

- Title
- Author(s)
- Sponsorship (i.e., organization)
- Date
- Classification (i.e., discipline)
- Subject (i.e., keywords)
- Citation
- Description
- Publisher
- Type (e.g., technical report, article, audio recording)
- URI

Storing accurate and useful metadata makes searching for relevant documents simple. The data files that are organized into related sets are then grouped into items. An item is an "archival atom" consisting of grouped, related content and associated descriptions (metadata). An item's exposed metadata is indexed for browsing and searching. Items are organized into collections of logically related material. The highest level of DSpace content hierarchy is a community, a collection of items. A community corresponds to parts of the organization implementing the DSpace such as a department, lab, research center, etc. The end user accesses the files in DSpace via a web interface. Once an item is located, Web-native files can be displayed in a Web browser while other formats can be downloaded and opened with suitable software. The Trinity River Basin system can be accessed at: http://repositories.lib.utexas.edu/handle/2152/4029.

5.7 THE TEXAS WATER DIGITAL LIBRARY

A pilot project has been initiated in Texas to implement and evaluate a statewide digital library for water resources information and issues. The goal of the Texas Water Digital Library (TWDL) is to be a centralized, online location for the research and works of university and other water resource entities in Texas, effectively federating water research currently housed at several universities across the state (https://repositories.tdl.org/twdl-ir/) (Figure 50). The TWDL began a partnership between water researchers and digital library professionals and librarians. Founding members include Dr. John Leggett of the Texas A&M University Libraries and Mark McFarland of the University of Texas Libraries, co-directors of the Texas Digital Library (TDL), plus the Directors of the Texas Water Resources Institute (TWRI) at Texas A&M, the Center for Research in Water Resources (CRWR) at UT-Austin, and the Water Research Center (WRC) at Texas Tech University.

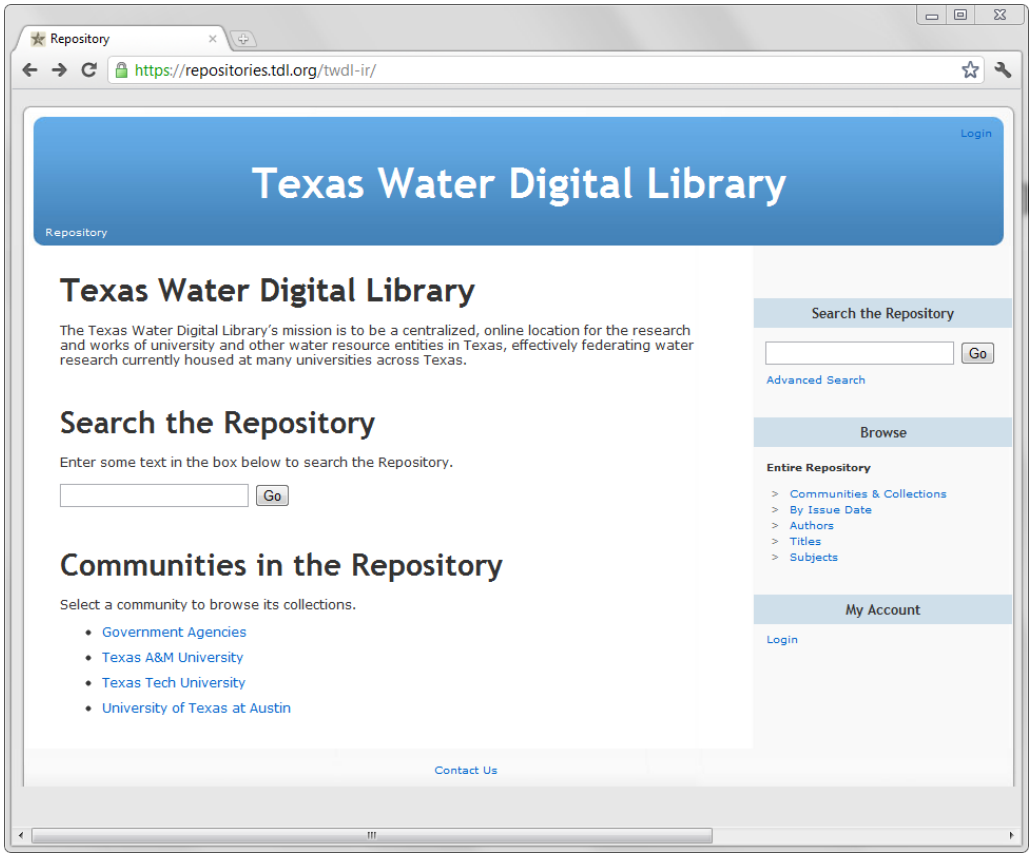


Figure 50. The Texas Water Digital Library homepage, https://repositories.tdl.org/twdl-ir/.

One featured digital repository of the TDL is that of Texas A&M University (http://repository.tamu.edu/), "a digital service that collects, preserves, and distributes the scholarly output of the university. The repository facilitates open access scholarly communication while preserving the scholarly legacy of Texas A&M faculty." Similarly, the featured repository of the University of Texas at Austin (http://repositories.lib.utexas.edu/) has the stated purpose "to collect, record, provide access to, and archive the scholarly and research works of the University of Texas at

Austin, as well as works that reflect the intellectual and service environment of the campus." Both of these repositories, along with the TWDL, are built using the DSpace system.

A number of universities in Texas have successfully implemented university-wide digital libraries. Many of these same universities also have water research centers, and some of these research centers have leveraged the TDL infrastructure to create water-specific digital repositories. For example, TWRI at Texas A&M (http://repository.tamu.edu/handle/1969.1/6061), WRC at Texas Tech (http://esr.lib.ttu.edu/handle/2346/1732), and CRWR at UT-Austin (http://repositories.lib.utexas.edu/handle/2152/4028) currently host such research products as technical reports, project data, outreach publications, theses and dissertations.

One goal of the TWDL is to link written research products like articles and reports with their supporting analytic products – models and data. Two examples are offered of this: an academic study and a professional project.

A project was performed at UT-CRWR to develop and test a stream classification system for Texas to support analyses of environmental flows. The classification system is based on quantitative data for 18 distinguishing parameters encompassing watershed and stream channel processes from four disciplines: (1) Hydrology & Hydraulics, (2) Water Quality, (3) Geomorphology & Physical Processes, and (4) Climatology. The State of Texas was partitioned into five regions: East Texas, South-Central Texas, Lower Rio Grande Basin, West Texas, and North-Central Texas by 8-digit Hydrologic Unit Code (HUC) basins (Hersh and Maidment 2007).

A report describing the findings of the study is housed in the CRWR Digital Repository (http://repositories.lib.utexas.edu/handle/2152/7029) and its metadata has been harvested for inclusion in the TWDL (https://repositories.tdl.org/twdl-ir/handle/2152/7029). Should a researcher wish to further investigate the results of this study or use the data for further analysis, they can now get a data summary (https://goodnight.corral.tacc.utexas.edu/tacc/Collections/TWDL/StreamClass/StreamClassificationDataSummary_CRWR.doc), the project tabular data (https://goodnight.corral.tacc.utexas.edu/tacc/Collections/TWDL/StreamClass/StreamClassificationData_CRWR.xls), and the project geospatial data (https://goodnight.corral.tacc.utexas.edu/tacc/Collections/TWDL/StreamClass/StreamClassification_CRWR.zip) from the TWDL. These data are housed in the iRODS data system on the Corral Server at TACC and the URLs are linked via the TWDL metadata.

But TWDL does not have to be limited to academic research endeavors – it can also support public-private partnerships and even private analyses (providing that necessary permissions have been granted for intellectual property access). An example of such a partnership, an updated flood study, is presented here as representative of a common water resources project typically undertaken by the engineering consulting community in Texas.

5.8 WATER INFORMATION SYSTEM OF SYSTEMS

As part of this study, Hydrologic Information Systems were implemented as part of two disparate case studies – freshwater rivers in Texas and the offshore shelf ecosystem in the Alaskan Arctic. Although the environmental conditions and the array of species present couldn't be more dissimilar in these two case studies, the observations data resulting from each of these two studies is remarkably similar in character and structure. As such, a common framework for water information management for terrestrial and marine systems is emerging.

If the information derived from a study was just observations data, then a relational database in an existing Hydrologic Information System alone would be sufficient. If the information was only geographic in nature, a geodatabase in a Geographic Information System would be sufficient. If the information was only articles, reports, and documents, a Digital Library would be sufficient. However, modern studies of the water environment are often multi- and interdisciplinary and highly complex in nature, encompassing some or all of the information types discussed above.

Given all of these considerations, it is recognized there is no single tool which will adequately fit the project information management needs, but rather some combination of tools is required – a "system of systems." This system of systems likely includes multiple means of data storage and multiple avenues of data access aggregated into a shared data portal. It likely includes water web services, maps, documents, and databases. As such, it is envisioned that the next generation of Hydrologic Information

Systems will be comprised of three component systems and will thus actually be a Water Information System of Systems (Figure 51, Figure 52):

1. Hydrologic Information Systems (HIS) for observations data,
2. Geographic Information Systems (GIS) for geographic data, and
3. Digital Libraries for digital assets (documents, images, videos).

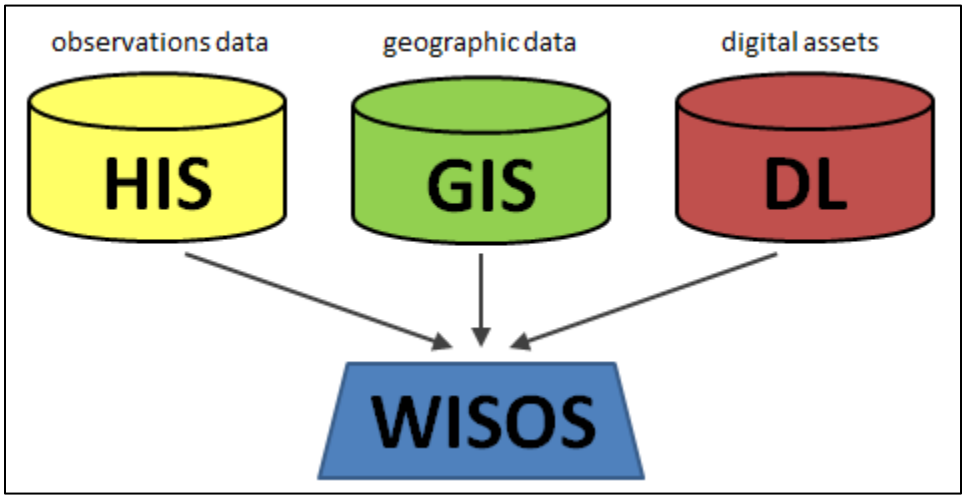


Figure 51. Water Information System of Systems schematic representation.

Note that the "system of systems" proposed here is a system of *information* systems. As such, the Water Information System of Systems differs from the Global Earth Observing System of Systems (GEOSS), which is proposed as a system of *observing* systems (i.e., from satellites and other data collection networks). Although the general concepts of information synthesis and integration are applicable in both cases, the scope and challenges associated with the development of each of these "system of systems" varies greatly.

In many instances, comprehensive project information management requires the multi-system approach detailed in Figure 52. The traditional starting point for information management is data, often formatted in individual files, stored on a personal computer, shared via email, and so forth. This level of data management is cheap and easy, but offers limited searchability and minimal (if any) organization.

Relational databases offer a structured and organized data storage solution and support for queries and more advanced analysis, but require additional expertise and resources for successful application. Some generic database programs in wide use are Microsoft Access, Oracle, and Microsoft SQL Server. Relational databases have been customized for use in the hydrologic sciences and are the central component of a Hydrologic Information System.

If the project information to be managed has a spatial component (such as a county, watershed, stream, well, or sampling location), data themes can be used to store the geographic representation of the observations data, commonly managed inside a Geographic Information System. Historically, proprietary software such as ESRI ArcGIS has been used for GIS databases, but spatial data are increasingly stored and shared on the web via ArcGIS Online, Google Maps, Google Earth, and portals like Geo.Data.gov (http://geo.data.gov/geoportal/catalog/main/home.page). Data themes offer place-based awareness and can greatly aid in technical communications but are not as robust of a metadata storage solution as are relational databases.

If the project information to be managed includes journal articles, project reports, photos, video, or other digital assets, a digital library offers the appropriate storage

solution for both the bitstream itself (i.e., the article or report) and for the associated metadata. Digital libraries are growing in popularity and offer a searchable interface for access to products of 'higher knowledge content' – the results of analysis and interpretation, not just raw data. However, the articles and reports must somehow be linked back to the data itself to support reanalysis.

	Data	Databases	Data Themes	Digital Objects
Definition	individual files	structured and organized data	geographic representation of observations data	articles, documents, images
Format	digital (csv, xls, txt, etc) or paper	SQL, Oracle, Access	shp, feature class, kmz	many; often pdf or image files (jpg, tif, etc)
Curator	individual researchers/ scientists	data managers	GIS analysts	digital librarians
Organizing	worksheets, files	HIS	GIS	digital repositories
Storing	file folders	ODM	geodatabase	metadata plus bitstream
Archiving	external media, enterprise file system, cabinets/ boxes	HIS Central, HydroServer	GeoPlatform, geo.data.gov	Digital Libraries (eg: UTDR, TDL)
Communicating	email, flashdrive, ftp	WaterOneFlow web services	OGC services, map services	bitstream
Advantages	easy, familiar, inexpensive	supports queries and analysis; well-documented metadata	place-based awareness; communications	includes higher knowledge-content products such as interpretation and analysis
Disadvantages	limitedly searchable; susceptible to changes in software, hardware, staffing turnover, etc	complex data loading requirements; data manager necessary	loss of some data fidelity/ quality (methods, qualifiers, etc)	does not include access to the raw data for new or re-analysis

Figure 52. Proposed Water Information System of Systems.

5.9 IMPACT TO-DATE

Google Analytics has been installed on the http://efis.crwr.utexas.edu/ website to anonymously gather data on the site's audience and traffic sources (Figure 53). Since its launch in December 2009, the EFIS website has experienced 1526 visits from 1024 unique visitors for a total of 4616 page views (Table 17). The average visitor views 3.02 pages and spends 3:14 on the site per visit. Visitors hail from 32 countries and 43 US states, with 35% visiting the site more than once and 10% visiting 10 or more times!

Table 17. Summary of audience and usage statistics for EFIS, COMIDA CAB, and Texas seagrass data portals.

Data Portal	# of Visits	# Unique Visitors	Time (Years)
EFIS	1526	1024	< 3
COMIDA CAB	2200	1109	< 2.5
Texas seagrass	799	452	< 2
TOTAL	**4525**	**2585**	

Besides EFIS, Google Analytics has also been installed on the COMIDA CAB project site (http://comidacab.org/) and a Texas seagrass monitoring project site (http://texasseagrass.org/). The seagrass site includes details on a statewide monitoring program on seagrass presence, condition, and associated environmental factors along

with documents, presentations, sampling design, maps, geostatistical results, and the complete project database. (Wilson and Dunton 2012)

For these three data portals, EFIS has received 1526 visits from 1024 unique visitors in less than 3 years; COMIDA CAB has received 2200 visits from 1109 unique visitors in less than 2-1/2 years; and Texas Seagrass has received 799 visits from 452 unique visitors in less than 2 years. In total, these three information systems have 4525 visits from 2585 unique visitors who have collectively spent 189 hours on the sites – nearly 7.9 days!

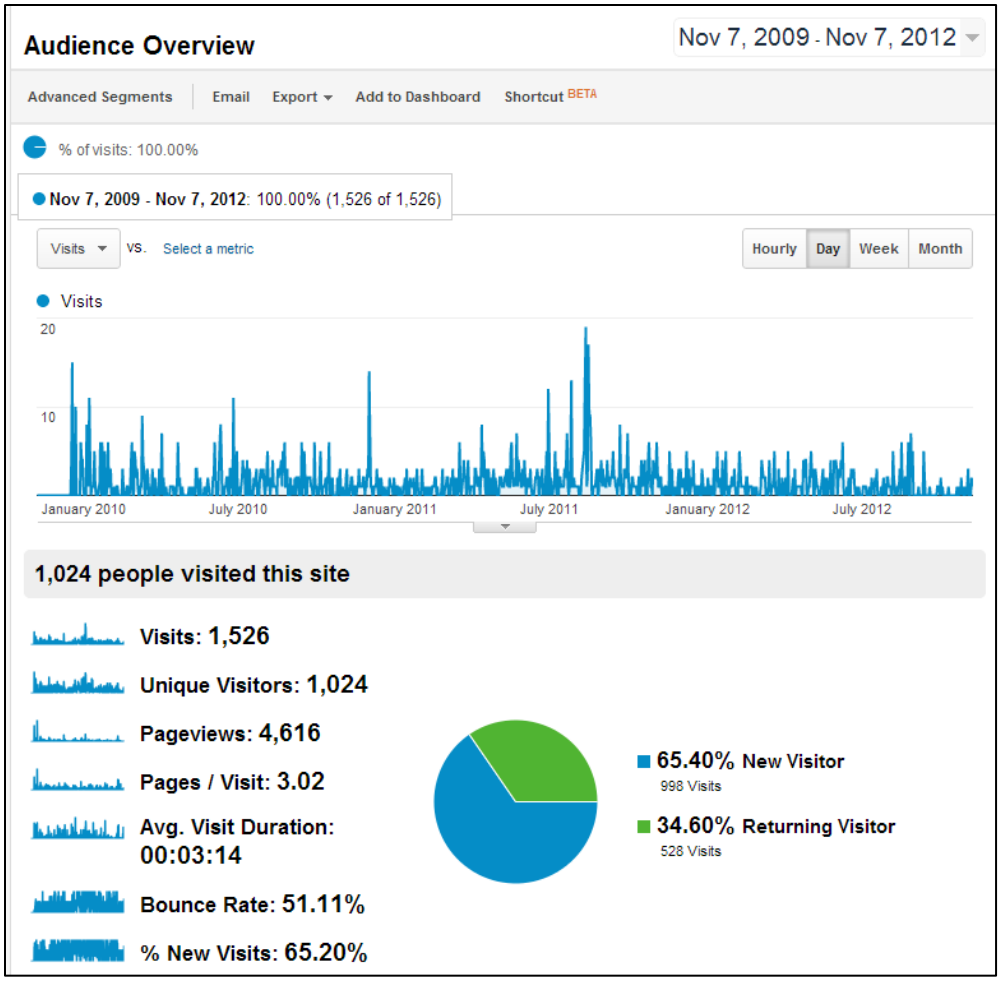


Figure 53. Overview of the EFIS website audience, November 7, 2009 to November 7, 2012, as obtained via the Google Analytics tools.

5.10 CONCLUSIONS

An example case study of an HIS implementation was presented in this chapter – the Texas Environmental Flows Information System (EFIS) – which aggregates and makes available information on hydrology, biology, water quality, and geomorphology in

support of the determination of environmental flow needs in Texas bays and basins for use by practitioners and stakeholders involved in that process across the state. In addition, a specific, detailed case study of invasive fish species in the Sabine River in Texas was presented as an example of the type of analysis which can be accomplished with access to robust biological databases.

It is hoped that the Texas Environmental Flows Information System (EFIS) might be used to provide: (1) rapid low-cost data integration, (2) improved data access by the public, and (3) support for the analysis and determination of environmental flow needs. EFIS represents the integration of the physical, chemical, and biological information for rivers and streams in a consistent and accessible manner in one system in one place.

In this chapter, a complete, real-world implementation of an expanded Hydrologic Information System was presented, inclusive of biological observations data, geographic data, tools, and documents, for a highly multi- and interdisciplinary arena – the field of environmental flows. As compared to the Arctic case study presented earlier, the Texas environmental flows case study includes information provided by many more partners hailing from public, private, university, and not-for-profit organizations, and has as its target audience a much more diverse community of stakeholders and practitioners working together in the determination of environmental flows needs in Texas.

Furthermore, it was recognized that no single tool is sufficient for complete project information management, but rather some combination of tools. As such, a Water Information System of Systems was introduced which consists of a Hydrologic Information System, a Geographic Information System, and a Digital Library. This

system of systems likely includes multiple means of data storage and multiple avenues of data access aggregated into a shared data portal. It likely includes water web services, maps, documents, and databases. And it is envisioned that the next generation of Hydrologic Information Systems will be comprised of these three component systems and will thus actually be a Water Information System of Systems.

Chapter 6: Conclusions

6.1. Addressing the Research Questions

Three research questions were posed at the beginning of this dissertation which served to guide the work discussed herein. Through the chapters provided here, the research questions have been addressed in the following manner:

1. ***How can existing Hydrologic Information Systems which focus largely on physical and chemical data be made more robust to accommodate biological data?***

This research questions was addressed via an examination of the issues associated with biological data integration, via the conceptualization of a data model for biological information, via an elaboration of use cases and scenarios, and via improvements and expansions to the information model currently in use for Hydrologic Information Systems.

This work has shown that existing Hydrologic Information Systems can accommodate biological data with some modifications. Of particular importance is the introduction of the 4-D data cube which considers both species and traits in place of the traditional 'variables' axis of the 3-D data cube. A corresponding BioODM data model

was conceived and developed for the four-dimensional biological observations data of species plus trait. Because of this treatment, databases can be queried by species ("I want to know everything about Guadalupe Bass (*Micropterus treculii)* in the Blanco River.") and also by trait ("What is the relative abundance (the trait) of Bering Flounder (*Hippoglossoides robustus)* observed in the Beaufort Sea?" or "What is the average length (a statistic calculated on the trait "length") of Bering Flounder observed in the Beaufort Sea?") Additionally, the reworked and expanded biological domain of the CUAHSI HIS ontology allows for the incorporation of biological observations in the same system and on the same plane as physical and chemical observations.

2. ***How can existing Hydrologic Information Systems which focus largely on observations of the terrestrial water environment be made more robust to accommodate oceanographic data?***

This research questions was addressed via an analysis of the nature of oceanographic observations data and a comparison with the nature of terrestrial aquatic data, via discussion of observing the ocean environment, organizing and storing oceans data, and communicating the results.

This work has shown that existing Hydrologic Information Systems can accommodate oceanographic data with some other modifications. What was accomplished via this case study was the adaptation of the CUAHSI Observations Data Model for application with physical, chemical, and biological oceanographic data – a new

extension of the CUAHSI Hydrologic Information System. Furthermore, the need for accommodating biological observations of the ocean environment in a cohesive project database drove further refinement of the BioODM data model, which was also amended to include better source tracking for the novel chain-of-custody approach introduced here for more robust data quality control, accountability, transparency, and persistence. In this sense, a complete, real-world implementation of an expanded Hydrologic Information System is presented, inclusive of biological and oceanographic data, for a multidisciplinary academic study.

3. Is there a common framework for water information management for terrestrial and marine systems?

This research questions was addressed via a detailed literature and technology review of existing tools and systems, via an investigation and assessment of digital library technologies, and via the introduction of a more robust 'system of systems' for water information which can accommodate geographic data via inclusion of a Geographic Information System, observations data via inclusion of a traditional Hydrologic Information System, and documents and other digital assets via inclusion of a Digital Library (as opposed to existing systems which can only accommodate point observations data).

This work has shown that many common elements exist between terrestrial and marine systems and a common framework has emerged and is emerging as a result. This

is a somewhat surprising result given the significant structural and functional differences which exist between freshwater and marine ecosystems. Also, this work has shown that Digital Libraries play an important role in the Water Information System of Systems of the future, envisioned to be a composite system of a Hydrologic Information System for observations data, and Geographic Information System for geospatial data, and a Digital Library for documents and other digital assets.

6.2. CONTRIBUTIONS TO SCIENCE AND TECHNOLOGY

By addressing the research questions posed above, this research contributes to the current state of science and engineering, particularly in that this work extends current Hydrologic Information System capabilities by providing additional capacity and flexibility for marine physical and chemical observations data and for freshwater and marine biological observations data. The case studies presented herein led to the development of a new four-dimensional data cube to accommodate biological observations data with axes of space, time, species, and trait. Collectively, the information systems and data portals presented here offer: (1) improved access to biological data and information for the freshwater environment; (2) improved access to oceanographic data and information for the marine environment; and (3) improved data discovery and data storage methodologies for freshwater and marine environments.

The BioODM data model presented here offers improved biological data management and represents progress toward the ultimate goal of synthesis and

integration of the physical, chemical, and biological elements of the water environment. Likewise, the two case studies conducted here have led to an improved understanding of the common elements and common framework of terrestrial and marine water information systems. The chain-of-custody approach introduced here represents better source tracking and thus a step toward more complete and more seamless knowledge management. The Water Information System of Systems introduced here is a vision for the next generation of Hydrologic Information Systems in that it includes far greater adaptability for multiple types of information – geographic data stored in a geodatabase in a Geographic Information System, observational data stored in a relational database in a Hydrologic Information System, and documents and other digital assets stored in a digital repository in a Digital Library.

This research contributes to the advancement of the CUAHSI Hydrologic Information System in a number of specific ways as well. This research improves our understanding of how to deal with collections of biological data stored alongside sensor-based physical data. The new 4-D data cube for biological observations data with axes of space, time, species, and trait represents an improved model for data storage, and the reworking and expansion of the biological domain ontology represents improved semantic mediation and a more robust data dictionary for biological observations.

6.3. RECOMMENDATIONS FOR FUTURE WORK

While significant progress in hydroinformatics has been made in data storage and communication, much work remains to be done in data access (Tarboton et al. 2010). For example, how do users interact with information systems? What tools and applications do they need? One promising avenue of exploration is the use of web geoprocessing services to provide online analytical capabilities. The Calculator for Low Flows (CaLF) tool uses web services to automatically access USGS streamflow data to assess low flows and environmental flows; this approach can be expanded for a broader assessment of flow regimes and could be made geospatially-explicit – for different freshwater ecoregions, different state requirements, and different stream types. And how can new user communities and data types be incorporated?

A session was convened at the 2011 AGU Fall Meeting discussing "data scientists," an emerging role which defines those who can effectively communicate with both domain specialists (scientists, ecologists, etc) and data managers (database experts, IT specialists, etc). The role of the data scientist needs to continue to be defined and these professionals need to be better leveraged as a means to bridge data managers and domain specialists.

Better education and training in informatics is needed to best prepare data scientists, and all scientists and engineers who work with data would benefit from this training as well. Many accredited undergraduate civil engineering programs in the United States require coursework in surveying, drafting, and basic computer

programming; what about informatics, analytics, data mining, and advanced statistics? The expanding role and presence of data-intensive science is driven, to some degree, by new and evolving data management requirements by the National Science Foundation (NSF) and others. Better training in informatics and analytics will better prepare the next generation of scientists and engineers to be completive in the new global economy.

Although the original CUAHSI HIS project has completed, its ground-breaking work in hydroinformatics could be expanded in a number of important directions in order to increase its utility for a broader audience (Figure 54).

	CUAHSI HIS focuses on...	...but could be expanded to include
WHO	Primarily sensor networks	Individual researchers and discrete studies
WHO	Data provider information	Traceable chain-of-custody information
WHAT	Physical and chemical data	Biological data
WHAT	Observations data stored in a relational database	Geographic data in a geodatabase and documents in a digital library
WHERE	Point observations	Lines (e.g. reaches), polygons (watersheds) and rasters (gridded data)
WHERE	Rivers, bays and estuaries	Oceans
WHEN	Time series data	Instantaneous and irregular data
HOW	Sampling method	Sampling effort exerted (e.g. Catch Per Unit Effort)

Figure 54. Possible directions for the expansion of current CUAHSI Hydrologic Information System focus areas.

Similarly, an improved global geospatial model for riverine systems would enhance our understanding of the lotic environment and would facilitate improved data management – it's time to move beyond "blue lines." Ideally, this geospatial model would incorporate four-dimensional data, with physical data georeferenced in both rectangular coordinates as {x,y,z} and also as curvilinear (a.k.a. fluvial) coordinates as {s,n,z} (Table 18), both of which are reflective of the dynamic nature of fluvial systems with respect to geomorphology and hydrology. Using fluvial coordinates simplifies and enhances the process of linear referencing and stream addressing. By moving to such a 4-D model, scientist and engineers can better answer such questions as: Was a Dissolved Oxygen sample taken near the surface? In the water column? At the sediment-water interface? In the pore space?

Table 18. Three-dimensional curvilinear coordinate system for stream network linear referencing.

s	Relative linear-referenced stream address (analogous to River Mile)
n	Width offset at cross-section (from bank, thalweg, etc)
z	Depth offset (from water surface, channel bed, etc)

It is also important to note that hydrologic sciences exist *outside* just the river channel. In the sense used in this research, hydrology includes all aspects of the global

water cycle – on the land surface, in the oceans, below the ground surface, and in the atmosphere. As new research increasingly focuses on hydrologic processes at the interface of various systems, hydrologic information systems must advance accordingly in order to support these research frontiers.

In conclusion, it is recognized that the collective value of Long Tail data is enormous. As such, it is hoped that the tools and systems presented herein serve to advance the field of hydroinformatics, especially with respect to biological and oceanographic observations of the water environment, in order to help harness the collective power of the Long Tail.

Glossary

7Q1	Seven-day average, one-year recurrence interval discharge
7Q2	Seven-day average, two-year recurrence interval discharge
ADCP	Acoustic Doppler Current Profiler
AGU	American Geophysical Union
API	Application Programming Interface
API	American Petroleum Institute
ARS	Agricultural Research Service
ASCE	American Society of Civil Engineers
ASCII	American Standard Code for Information Interchange
AWRIS	Australian Water Resources Information System
BBEST	Bay and Basin Expert Science Teams
BioODM	Biological Observations Data Model
BOEM	Bureau of Ocean Energy Management
CAB	Chemical and Benthos
CaLF	Calculator for Low Flows
CFR	Code of Federal Regulations
CFS	Cubic feet per second
CI-WATER	Cyberinfrastructure-Water
COMIDA	Chukchi Sea Offshore Monitoring in Drilling Area
CONABIO	Comisión Nacional para el Conocimiento y Uso de la Biodiversidad of Mexico
CRWR	Center for Research in Water Resources
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSV	Comma-separated value
CTD	Conductivity, temperature, and depth

CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science, Inc.
DAR	Drainage Area Ratio
DCMI	Dublin Core Metadata Initiative
DOI	United States Department of Interior
EDAS	Ecological Data Application System
EFIS	Environmental Flows Information System
EMAP	Environmental Monitoring and Assessment Program
EML	Ecological Markup Language
EOL	Earth Observing Laboratory
EPA	United States Environmental Protection Agency
ETL	Extract-Transform-Load
FBIS	Freshwater Biodata Information System
FGDC	Federal Geographic Data Committee
G8	Group of Eight
GBIF	Global Biodiversity Information Facility
GEO	Group on Earth Observing
GEOSS	Global Earth Observing System of Systems
GIS	Geographic Information System
GRTS	General Randomized Tessellation Stratified design
HEC-RAS	Hydrologic Engineering Center – River Analysis System
HEFR	Hydrology-Based Environmental Flow Regime
HIS	Hydrologic Information System
HTML	HyperText Markup Language
HUC	Hydrologic Unit Code
IBI	Index of Biological Integrity
IHA	Indicators of Hydrologic Alteration
IHE	Institute for Hydraulic and Environmental Engineering
IPCC	Intergovernmental Panel on Climate Change

iRODS	Integrated Rule-Oriented Data System
IT	Information Technology
ISO	International Organization for Standardization
ITIS	Integrated Taxonomic Information System
IUCN	International Union for Conservation of Nature
LDAP	Lightweight Directory Access Protocol
LDCurve	Load Duration Curve tool
LTER	Long-Term Ecological Research
MIT	Massachusetts Institute of Technology
MMS	Minerals Management Service
NBII	National Biological Information Infrastructure
NCAR	National Center for Atmospheric Research
NCEAS	National Center for Ecological Analysis and Synthesis
NEON	National Ecological Observation Network
NGDC	National Geophysical Data Center
NIWA	National Institute of Water and Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
NPS	National Park Service
NRCS	Natural Resources Conservation Service
NSF	National Science Foundation
NTL	North-Temperate Lakes
NWIS	National Water Information System
NWS	National Weather Service
OBIS-SEAMAP	Ocean Biogeographic Information System – Spatial Ecological Analysis of Marine Megavertebrate Animal Populations
ODM	Observations Data Model
OGC	Open Geospatial Consortium
PAR	Photosynthetically Active Radiation

PI	Principal Investigator
POC	Particulate Organic Carbon
POM	Polycyclic Organic Matter
QA/QC	Quality Assurance/ Quality Control
R/V	Research Vessel
RENCI	Renaissance Computing Institute at UNC Chapel Hill
REST	Representational State Transfer
ROV	Remotely-Operated Vehicle
SB3	Senate Bill 3
SBI	Western Arctic Shelf-Basin Interactions project
SOA	Services-Oriented Architecture
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SSIS	SQL/Server Integration Services
STORET	Storage and Retrieval
SWQM	Surface Water Quality Monitoring
SWQMIS	Surface Water Quality Monitoring Information System
TACC	Texas Advanced Computing Center
TCEQ	Texas Commission on Environmental Quality
TCOON	Texas Coastal Ocean Observation Network
TDL	Texas Digital Library
TNRIS	Texas Natural Resources Information System
TOC	Total Organic Carbon
TPWD	Texas Parks and Wildlife Department
TRACS	TCEQ Regulatory Activities Compliance Systems
TSS	Total Suspended Solids
TWDB	Texas Water Development Board
TWDL	Texas Water Digital Library
TWRI	Texas Water Resources Institute

UDDI	Universal Description, Discovery, and Integration
UML	Unified Modeling Language
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USDA	United States Department of Agriculture
USFWS	United States Fish and Wildlife Service
USGS	United States Geological Survey
UTC	Coordinated Universal Time
VBA	Visual Basic for Applications
WaterML	Water Markup Language
WCS	Web Coverage Service
WebDav	Web-based Distributed Authoring and Versioning
WFS	Web Feature Service
WMO	World Meteorological Organization
WMS	Web Map Service
WQX	Water Quality Exchange
WRC	Water Research Center
WSDL	Web Services Description Language
XML	Extensible Markup Language

References

Ackoff, R. (1989). "From Data to Wisdom." *Journal of Applied Systems Analysis*, 16:3-9.

Alavi, M. and Leidner, D.E. (2001). "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues." *MIS Quarterly*, 25(1):107-136.

Amante, C., Eakins, B. W. (2008). *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. National Geophysical Data Center, NOAA, U.S. Department of Commerce, Boulder, CO.

American Petroleum Institute. (2009). History of Northern Alaska Petroleum Development. http://www.api.org/aboutoilgas/sectors/explore/historyofnorthalaska.cfm

American Society of Civil Engineers. (2009). *2009 Report Card for America's Infrastructure*. ASCE, Washington, D.C. http://www.asce.org/reportcard/

Arc Advisory Group. (2010). *Geospatial Information Systems Worldwide Outlook: Five-Year Market Analysis and Technology Forecast through 2014.* http://ntlm.arcweb.com/study-brochures/Study_Geospatial-Info-Systems.pdf

Arctur, D. and Zeiler, M. (2004). *Designing Geodatabases: Case Studies in GIS Data Modeling.* ESRI Press, Redlands, California.

Atkins D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P., Wright, M.H. (2003). *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* http://www.nsf.gov/od/oci/reports/atkins.pdf

Atkins DE, et al. (2003). *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyber infrastructure.* www.communitytechnology.org/nsf_ci_report/

Baker, B. (2007). A conceptual framework for making knowledge actionable through capital formation. Dissertation, University of Maryland.

Beach, P. (2011). "Texas ranchers, farmers, seeing record losses in grip of drought, say, 'These are desperate times'." *Austin American-Statesman*. October 19, 2011. http://www.statesman.com/news/local/texas-ranchers-farmers-seeing-record-losses-in-grip-1917547.html

Bedig, A. and Couch, A. (2011). "Faceted Search for Hydrologic Data Discovery." CUAHSI Conference on Hydrologic Data and Information Systems, June 22-24, Logan, Utah. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&ved=0CDkQFjAB&url=http%3A%2F%2Fhis.cuahsi.org%2Fconference2011%2FPresentations%2F4.2.Bedig.pptx&ei=cQ2sUOOrKJLg2wXDmoCgBQ&usg=AFQjCNEBxg-AZfr8ijSIHRbrxvay5JQbHg&sig2=M375NinTTlWdopiTVWJtQA

Brisbane Declaration. (2007). "The Brisbane Declaration: Environmental Flows are Essential for Freshwater Ecosystem Health and Human Well-Being." 10th International Riversymposium and International Environmental Flows Conference. Brisbane, Australia, September 3-6, 2007. http://www.riversymposium.com/2004/index.php?element=2007BrisbaneDeclaration241007

Bisby F.A, Roskov Y.R, Orrell T.M, Nicolson D., Paglinawan L.E, et al., editors. (2010). "Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist." http://www.catalogueoflife.org/annual-checklist/2010.

Chong, F. and Carraro, G. (2006). "Architecture Strategies for Catching the Long Tail." Microsoft Developer Network (MSDN). http://msdn.microsoft.com/en-us/library/aa479069.aspx

Chow, V.T., Maidment, D.R. and Mays, L.W. (1988). *Applied Hydrology*. McGraw-Hill.

CI-WATER. (2011). "Researchers from Utah and Wyoming Join Forces to Understand Complex Water Problems Facing Western States." Press Release. http://ci-water.org/documents/press_release.pdf

Code of Federal Regulations. (2010). *Endangered and Threatened Wildlife and Plants; Designation of Critical Habitat for the Polar Bear (*Ursus maritimus*) in the United States*. Title 50, Part 17. U.S. Department of the Interior Fish and Wildlife Service.

Commonwealth Scientific and Industrial Research Organisation. (2009). "Australian Water Resources Information System." *CSIRO Land and Water Science Review*.

CUAHSI HIS. (2008). "Hydrologic Ontology for Discovery." http://his.cuahsi.org/ontologyfiles.html

CUAHSI HIS. (2010). "Development Wiki – HIS Glossary and Acronyms List." http://river.sdsc.edu/wiki/Default.aspx?Page=HIS%20Glossary&NS=&AspxAutoDetectCookieSupport=1#G_-_Q_thru_T_5

CUAHSI HIS. (2011). "CUAHSI Ontology." http://water.sdsc.edu/hiscentral/startree.aspx

CUAHSI HIS. (2012). "Services-Oriented Architecture." http://his.cuahsi.org

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., and Weerawarana, S. (2002). "Unraveling the Web services Web - An introduction to SOAP, WSDL, and UDDI." *IEEE Internet Computing*, 86-93.

DSpace. (2009). "DSpace." Developed by the MIT Libraries and Hewlett-Packard, Inc. http://www.dspace.org/

Dublin Core Metadata Initiative. (2009). "DCMI Specifications." http://dublincore.org/specifications/

Edwards, J.L., Lane, M.A., and Nielsen, E.S. (2000). "Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop." *Science*, 289:2312-2314.

EarthCube. (2012). "Welcome to EarthCube, transforming geosciences research for the 21st century." http://earthcube.ning.com/

Elias, T. (2011). *Learning Analytics: Definitions, Processes, and Potential*.

Environmental Data Services. (2008). "ISEMP Aquatic Resources Metadata Framework." Prepared for the Integrated Status and Effectiveness Monitoring Program of the Northwest Fisheries Science Center. Portland, Oregon.

Erl, T. (2004). *Service-Oriented Architecture.* Upper Saddle River: Prentice Hall PTR.

Erl, T. (2005). *Service-Oriented Architecture (SOA): Concepts, Technology, and Design.* Upper Saddle River: Prentice Hall PTR.

ESRI (2010). ArcGIS Online http://www.arcgisonline.com

ESRI (2010b). ArcGIS Resource Center: Data Models. http://resources.arcgis.com/content/data-models

Fannin, B. (2011). "Texas agricultural drought losses reach record $5.2 billion." AgriLife Today. Texas AgriLife Extension Service, Texas A&M University. August 17, 2011. http://agrilife.org/today/2011/08/17/texas-agricultural-drought-losses-reach-record-5-2-billion/

Federal Geographic Data Committee. (2009). "Geospatial Metadata." http://www.fgdc.gov/metadata

Foresman, T.W., Ed. (1998). *History of Geographic Information Systems: Perspectives from the Pioneers*. Prentice Hall PTR.

Froese, R. and Pauly, D. Ed. (2010). "FishBase, version (01/2010)." http://www.fishbase.org

Global Biodiversity Information Facility. (2010). http://www.gbif.org/

Goodall, J., Horsburgh, J., Whiteaker, T., Maidment, D., and Zaslavsky, I. (2008). "A first approach to web services for the National Water Information System." *Environmental Modelling & Software*, (23) 404-411.

Greenstein, D. and Thorin, S.E. (2002). *The Digital Library: A Biography*. Council on Library and Information Resources.

Group on Earth Observations. (2012). The Global Earth Observing System of Systems. http://www.earthobservations.org/geoss.shtml

Gruber, T. (1995) "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." *International Journal Human-Computer Studies*, 43(5-6):907-928.

Halpin, P.N., Read, A.J., Best, B.D., Hyrenbach, K.D., Fujioka, E., Coyne, M.S., Crowder, L.B., Freeman, S.A., and Spoerri, C. (2006). "OBIS-SEAMAP: developing a Biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles." *Mar. Ecol. Prog. Ser*., 316:239-246.

Hersh, E.S. (2007). *New tools for interdisciplinary river research in Texas: flow regime assessment and stream classification*. Thesis. The University of Texas at Austin.

Hersh, E.S. and Maidment, D.R. (2007). *An integrated stream classification system for Texas*. CRWR Online Report 07-02. Center for Research in Water Resources, University of Texas at Austin. http://www.crwr.utexas.edu/reports/2007/rpt07-2.shtml

Hersh, E.S., Marney, K.A., and Maidment, D.R. (2008). *Trinity River Basin Environmental Flows Information Collective*. CRWR Online Report 08-08. http://www.crwr.utexas.edu/reports/2008/rpt08-8.shtml

Horsburgh, J., Tarboton, D., Maidment, D., and Zaslavsky, I. (2008). "A relational model for environmental and water resources data." *Water Resources Research*, (44).

Hubbs C., Edwards, R.J., and Garrett, G.P. (1991). "An Annotated Checklist of the Freshwater Fishes of Texas, with Keys to Identification of Species." Supplement to the *Texas Journal of Science*, 43(4).

Integrated Taxonomic Information System. (2012). "About ITIS." http://www.itis.gov/info.html

International Organization for Standardization. (2003). *ISO 19115 Geographic Information - Metadata*. Geneva, Switzerland. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020

International Union for Conservation of Nature. (2011). *The IUCN Red List of Threatened Species*. http://www.iucnredlist.org/

Jacobson, I., Christerson, M., and Jonsson, P. (1992). *Object-oriented software engineering: a use case driven approach*. ACM Press, New York.

Jelks, H.L., et al. (2008). "Conservation status of imperiled North American freshwater and diadromous fishes." *Fisheries* 33(8):372-407.

Johnson, S. L., Whiteaker, T., and Maidment, D. (2008). A Tool for Automated Load Duration Curve Creation. *Journal of the American Water Resources Association*, 45(3):654-663.

Johnson, S.L. (2009) *A general method for modeling coastal water pollutant loadings*. Dissertation. The University of Texas at Austin.

Josuttis, N.M. (2007). *SOA in practice – the art of distributed system design*, O'Reilly Press, Sebastopol, California.

Kainerstorfer, C. and Perkins, H. (2009). "Intro to DSpace." http://www.tdl.org/wp-content/uploads/2009/08/IntroToDSpace_081920091.pdf

Karr, J.R. (1981). "Assessment of biotic integrity using fish communities." *Fisheries*, 6:21-27.

Kumar, P. et al. (2006). *Hydroinformatics: data integrative approaches in computation, analysis, and modeling*. CRC Press, Taylor and Francis Group, Boca Raton, Florida.

Linnaeus, Carolus. (1735). *Systema naturæ, sive regna tria naturæ systematice proposita per classes, ordines, genera, & species* (A general system of nature, of the three kingdoms of nature, and systematically proposed to classes, orders, genera, and species). Lugduni Batavorum.

Linnaeus, Carolus. (1758). *Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* (A general system of nature: through the three grand kingdoms of animals, vegetables, and minerals, systematically divided into their several classes, orders, genera, species, and varieties) (10th edition). Stockholm: Laurentius Salvius.

Linam, G.W., Kleinsasser, L.J, and Mayes, K.B. (2002). "Regionalization of the Index of Biotic Integrity for Texas Streams." River Studies Report No. 17, Texas Parks and Wildlife Department. http://www.tpwd.state.tx.us/publications/pwdpubs/media/pwd_rp_t3200_1086.pdf

Long Term Ecological Research Network. (2010). "The US Long Term Ecological Research Network." http://www.lternet.edu/

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). "An ontology for describing and synthesizing ecological observation data." *Ecological Informatics*, 2:279-296.

Maidment D.R. (2002). *Arc Hydro: GIS for Water Resources*. ESRI Press, Redlands, California.

Maidment, D.R., Ed. (1993). *Handbook of Hydrology*. McGraw-Hill Professional.

Maidment, D.R., Ed. (2008). "CUAHSI Hydrologic Information System: Overview of Version 1.1", http://his.cuahsi.org/documents/hisoverview.pdf

Maidment, D.R., Ed. (2009). "CUAHSI HYDROLOGIC INFORMATION SYSTEM: 2009 Status Report." http://his.cuahsi.org/documents/HISOverview2009.pdf.

McCall, R.A. and May, R.M. (1995). "More than a Seafood Platter." *Nature*, 376:735.

Milly, P.C.D et al. (2008). "Stationarity is Dead: Whither Water Management?" *Science* 319:573-574. http://wwwpaztcn.wr.usgs.gov/julio_pdf/milly_et_al.pdf

Minerals Management Service. (2008). Final Notice of Sale (FNOS) Outer Continental Shelf (OCS), Oil and Gas Lease Sale 193, Chukchi Sea. 73 Federal Register 1, pp. 209-213.

Minerals Management Service. (2008b). "MMS Chukchi Sea Lease Sale 193 Breaks Energy Records With $2.6 Billion in High Bids." News Release. http://www.alaska.boemre.gov/latenews/newsrel/News%20Releases%202008/News%20Release%20-%20193%20results%20_2_.pdf

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., and Worm, B. (2011). "How Many Species Are There on Earth and in the Ocean?" PLOS Biology, 9(8): 1-8. http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001127?imageURI=info:doi/10.1371/journal.pbio.1001127.t002

National Center for Atmospheric Research Earth Observing Laboratory. (2010). http://www.eol.ucar.edu/

National Center for Atmospheric Research Earth Observing Laboratory, (2011). http://www.eol.ucar.edu/

National Ecological Observation Network, Inc. (2010). About NEON. http://www.neoninc.org/

National Oceanographic Data Center. (2011). http://www.nodc.noaa.gov/

National Oceanographic Data Center. (2012). "NODC NetCDF Templates." http://www.nodc.noaa.gov/data/formats/netcdf/

National Weather Service. (2011). "Official U.S. Flood Loss Statistics." http://www.nws.noaa.gov/hic/flood_stats/

NatureServe. (2010). "NatureServe web services." http://services.natureserve.org/

North Temperate Lakes Long Term Ecological Research. (2008). http://lter.limnology.wisc.edu/

Open Geospatial Consortium. (2012a). http://www.opengeospatial.org/standards/netcdf

Open Geospatial Consortium. (2012b). "The OGC adopts WaterML 2.0 Hydrologic Time Series Encoding Standard." Press Release. September 20, 2012. http://www.opengeospatial.org/pressroom/pressrelease/1696

Pachauri, R.K and Reisinger, A., Eds. (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland. 104 pp. http://www.ipcc.ch/publications_and_data/ar4/syr/en/contents.html

Pielou, E.C. (1966). The measurement of diversity in different types of biological collections. *J. Theoretical Biology*, 13:131-144.

Rajasekar, A., M. Wan, R. Moore, W. Schroeder, 2006. A Prototype Rule-based Distributed Data Management System., HPDC workshop on "Next Generation Distributed Data Management," Paris, France.

Rajasekar, A., M. Wan, R. Moore, 2009. Event Processing in Policy Oriented Data Grids. Proceedings of Intelligent Event Processing AAAI Spring Symposium, Stanford, California, pp 61-66.

Renaissance Computing Institute (RENCI). (2012). "HydroShare aims to help scientists collaborate on water-related problems." Press Release. http://www.renci.org/news/releases/hydroshare-aims-to-help-scientists-collaborate-on-water-related-problems

Robertson, D. and de Winton, M. (2004). "FBIS: a new freshwater biodata information system." *Water and Atmosphere* 12(3):12-13.

Rowley, J. (2007). "The wisdom hierarchy: representations of the DIKW hierarchy." *Journal of Information Science*, 33(2): 163-180.

Ruggiero, M., McNiff, M., Olson, A., and Wheeler, B. (2005). *Strategic Plan for the U.S. Geological Survey Biological Informatics Program: 2005-2009*. U.S. Geological Survey, Biological Resources Discipline.

Sabine River Authority of Texas. (2007). *Baseline Fish Collections – Lower Sabine River Priority Instream Flow Study*. Final Report. Prepared by Sabine River Authority of Texas, Texas Commission on Environmental Quality, Texas Parks and Wildlife Department, and Texas Water Development Board under TWDB Research and Planning Fund Research Grant, Contract No. 0604830567.

Science Advisory Committee. (2010). *Discussion Paper: Moving from Instream Flow Regime Matrix Development to Environmental Flow Standard Recommendations*. Texas Senate Bill 3 Science Advisory Committee for Environmental Flows.

Science Advisory Committee. (2011). *Use of Hydrologic Data in the Development of Instream Flow Recommendations for the Environmental Flows Allocation Process and the Hydrology-Based Environmental Flow Regime (HEFR) Methodology, Third Edition.* Texas Senate Bill 3 Science Advisory Committee for Environmental Flows, Report #SAC-2011-01.fCaLF http://www.tceq.texas.gov/assets/public/permitting/watersupply/water_rights/eflows/hydrologicmethods06172011.pdf

Seppi, J. (2009). "ThemeViewer: An online viewer for geospatial space-time series." CE394K.3 – GIS in Water Resources term project. The University of Texas at Austin.

Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal. 27: 379-423 and 623-656

Shelf Basin Interaction. (2008). "SBI Data Archive." http://www.eol.ucar.edu/projects/sbi/cruise_summary_info.html

Siegler, M.G. (2011). "Google Maps For Mobile Crosses 200 Million Installs; In June It Will Surpass Desktop Usage." TechCrunch, May 25, 2011. http://techcrunch.com/2011/05/25/google-maps-for-mobile-stats/

Simpson, E. H. (1949) Measurement of diversity. Nature. 163:688.

Strassberg, G., Jones, N.L., and Maidment, D.R. (2011) *Arc Hydro Groundwater: GIS for Hydrogeology.* ESRI Press, Redlands, California.

Strassberg, G., Maidment, D.R. and Jones, N.L. (2007). "A Geographic Data Model for Representing Ground Water Systems." *Ground Water*, 45(4):515-518.

Tarboton, D.G. et al. (2010). *CUAHSI Hydrologic Information System 2010 Status Report*. http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf

Tetra Tech, Inc. (2000). *Ecological Data Application System (EDAS) – A User's Guide.*

Texas Commission on Environmental Quality. (2000). "Surface Water Quality Standards." Chapter 307 of the Texas Administrative Code.

Texas Digital Library. (2008). "Texas Digital Library." http://www.tdl.org.

The University of Texas at Austin Libraries. (2008). "UT DSpace Repository." http://repositories.lib.utexas.edu/

Thomas, C., Bonner, T.H., and Whiteside, B.G. (2007). *Freshwater Fishes of Texas*. College Station, Texas: Texas A&M University Press.

UNESCO-IHE Institute for Water Education. (2010). Hydroinformatics - Modelling and Information Systems for Water Management. http://www.unesco-ihe.org/Education/MSc-Programmes/MSc-in-Water-Science-and-Engineering/Hydroinformatics-Modelling-and-Information-Systems-for-Water-Management.

Unidata, 2012. http://www.unidata.ucar.edu/software/netcdf/

United Nations. (2011). *Millennium Development Goals*. http://www.un.org/millenniumgoals/

United States Environmental Protection Agency. (2010). "What is WQX?" http://www.epa.gov/storet/wqx/

United States Environmental Protection Agency. (2009). *National Water Quality Inventory: Report to Congress, 2004 Reporting Cycle*. USEPA, Washington, D.C. http://www.epa.gov/305b/

Vertessy, R. (2010). "Leveraging Australia's Water Information." Presentation to the Queensland Water Technical Reference Group, June 11, 2010, Brisbane.

White, D., Kimerling, J., Overton, S., 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartography and Geographic Information Systems*, 19(1): 5-21.

Whiteaker, T. (2009). "Thematic Dataset: Table Design." Unpublished CUAHSI HIS document. July 31, 2009.

Whiteaker, T., Maidment, D., Pothina, D., Seppi, J., Hersh, E., and Harrison, W. (2010). *Texas Hydrologic Information System*. AWRA 2010 Spring Specialty Conference: Orlando, Florida.

Whiteaker, T.L. (2008). *HydroObjects - Versions 1.1*. CUAHSI HIS. http://his.cuahsi.org/documents/HydroObjects_Software_Manual.pdf

Wilson, C.J. and Dunton, K.H. (2012). Assessment of seagrass habitat quality and plant physiological condition in Texas coastal waters. Final Report. http://texasseagrass.org/documents/Final%20Seagrass%20Report%20Tier%202%202012.pdf

Woese C., Kandler O., and Wheelis M. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–9.

Wood, L.F. and Williams, M. (2005). Vertebrate Field Zoology. Samford University. http://www4.samford.edu/schools/artsci/biology/zoology/vertzoo-05s/

Wright, D.J., Blongewicz, M.J., Halpin, P.N. and Breman, J. (2007). *Arc Marine: GIS for a Blue Planet*. ESRI Press, Redlands, California.

Zaslavsky, I., Valentine, D. and Whiteaker, T. (2007). "CUAHSI WaterML," OGC 07041r1, Open Geospatial Consortium Discussion Paper, http://portal.opengeospatial.org/files/?artifact_id=21743.

Vita

Eric Scott Hersh was born in September 1979 in Woodbridge, Connecticut. He graduated from The Hopkins School in New Haven in 1997 and then went on to Tufts University in Medford, Massachusetts, graduating in 2001 with dual majors of civil engineering and environmental studies. Eric worked as a water resource engineering consultant in Norwood, Massachusetts for GZA GeoEnvironmental, Inc. from 2001 through 2005 then completed a through-hike of the 2,175-mile Appalachian Trail from Georgia to Maine before beginning graduate studies at The University of Texas at Austin in August 2005. He received his Masters of Science degree in Environmental and Water Resource Engineering in August 2007 and his Professional Engineering license in Civil Engineering in 2009. He currently serves as the Research Program Coordinator at the University of Texas Environmental Science Institute and as a Lecturer in Environmental Science.

Permanent Email Address: ehersh@alumni.tufts.edu

This dissertation was typed by the author.